

A Comprehensive Review of Algorithms for Protein Structure Prediction and Docking

Giovanni Valentino*

Department of Biochemical and Environmental Engineering, Antonio García Cubas Pte #600 esq. Av. Tecnológico, Celaya 38010, Mexico

Introduction

Proteins, the molecular machines of life, are responsible for an array of biological functions, from catalyzing biochemical reactions to transmitting signals across cellular membranes. Their functions are intricately linked to their three-dimensional (3D) structures, making accurate protein structure prediction one of the most fundamental challenges in computational biology. Understanding a protein's structure is critical for drug design, disease understanding and biotechnology applications, among others. However, experimentally determining protein structures through techniques like X-ray crystallography, Nuclear Magnetic Resonance (NMR), or Cryo-Electron Microscopy (cryo-EM) is resource-intensive, time-consuming and not always feasible, especially for large proteins or membrane proteins. As a result, computational methods for Protein Structure Prediction (PSP) and protein-ligand docking have gained significant attention. These methods aim to predict a protein's 3D structure from its amino acid sequence and to model how proteins interact with small molecules, respectively. Protein-ligand docking, in particular, is crucial for drug discovery, as it predicts the binding affinity and orientation of ligands to their target proteins. This review delves into the state-of-the-art algorithms used for protein structure prediction and protein-ligand docking. We will discuss the theoretical foundations, different approaches, key techniques, challenges and recent advancements in the field [1].

Description

Protein structure prediction involves the determination of a protein's 3D structure from its linear amino acid sequence. This challenge stems from the fact that the protein folding problem is NP-hard, meaning there is no known algorithm that can efficiently predict protein structures for all cases. Over the years, several computational strategies have emerged, which can be broadly categorized into the following approaches: homology modeling, ab initio prediction and threading (also known as fold recognition). Homology modeling is based on the idea that protein sequences that share a significant level of similarity also share structural similarities. This approach involves identifying a homologous protein (template) whose structure is already known, aligning the target sequence with the template and then predicting the 3D structure of the target based on the template structure. Sequence alignment protein sequence is aligned with the sequence of a known protein structure. This alignment is crucial because even small differences in sequence can lead to significant changes in the 3D structure [2].

Energy-Based methods use force fields (mathematical models of atomic interactions) to calculate the energy of a given protein conformation. The goal

is to minimize the total energy to find the native structure. Fragment-based methods, such as those used in the Rosetta software suite, break the protein sequence into small fragments and search for the optimal combination of fragments that best fit the 3D structure. These methods combine the efficiency of sampling smaller pieces with the accuracy of using known structural motifs. Monte Carlo simulations are used to randomly sample the protein's conformational space, while molecular dynamics simulations simulate the physical movement of atoms over time. Both methods are computationally intensive but can provide high accuracy when properly applied. Ab initio prediction is computationally demanding and struggles with large proteins or proteins with disordered regions, where structural determination is particularly difficult. Threading, or fold recognition, is an intermediate method that is used when a close homolog is not available. It works by matching the target sequence against a library of known protein folds, rather than relying on sequence similarity alone. The algorithm "threads" the target sequence through a structural template, attempting to align it in such a way that it minimizes steric clashes and respects the protein's secondary structure [3].

Threading is more effective than pure sequence alignment when sequence similarity is low but structural similarity is still high. Modern threading methods integrate scoring functions that assess the quality of the structural alignment and can even predict the stability of the resulting model. Protein-ligand docking refers to the computational prediction of the binding mode and affinity of a ligand (typically a small molecule) to its target protein. This is a critical step in the drug discovery process, as it helps identify potential drug candidates and predicts their binding interactions before experimental validation. Protein-ligand docking methods are primarily divided into two categories: rigid docking and flexible docking. Rigid docking assumes that both the ligand and the receptor protein are inflexible during the docking process. In this approach, the protein's 3D structure is fixed and only the ligand is allowed to move, rotating and translating in space to find the optimal binding position. Ligand Conformation Generation is often provided as a single conformation, or a set of possible conformations, based on prior knowledge or rotatable bonds in the molecule. Algorithms such as the Fast Fourier Transform (FFT) and genetic algorithms search the docking space to identify the optimal orientation and position of the ligand. Scoring functions are used to evaluate the binding affinity of each ligand conformation by calculating the interaction energy between the ligand and the protein. These functions are critical for ranking different docking poses. Rigid docking is computationally efficient, but it is limited in its ability to accurately model the flexibility of the protein or the ligand, which can impact docking predictions [4].

Flexible docking methods allow both the protein and the ligand to undergo conformational changes during the docking process, making them more accurate for systems where flexibility plays a key role in the binding interaction. Flexible docking methods are typically more computationally demanding, as they require the exploration of larger conformational spaces. Both the ligand and the protein are allowed to adopt a range of conformations, which are then sampled during the docking process. The initial docking poses are refined using molecular dynamics or Monte Carlo simulations to further minimize the energy and predict more accurate binding modes. Flexible docking methods have become increasingly popular in drug discovery, as they provide a more realistic representation of ligand binding, but they come with a higher computational cost. Recent advancements in protein-ligand docking algorithms have focused on improving speed and accuracy. Several modern techniques have incorporated machine learning and artificial intelligence (AI) to improve the performance of docking simulations. For example, deep

*Address for Correspondence: Giovanni Valentino, Department of Biochemical and Environmental Engineering, Antonio García Cubas Pte #600 esq. Av. Tecnológico, Celaya 38010, Mexico; E-mail: valentinog@upgto.edu.mx

Copyright: © 2024 Valentino G. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Received: 26 July, 2024, Manuscript No. gjto-24-152498; Editor assigned: 29 July, 2024, Pre QC No. P-152498; Reviewed: 05 August, 2024, QC No. Q-152498; Revised: 12 August, 2024, Manuscript No. R-152498; Published: 19 August, 2024, DOI: 10.37421/2229-8711.2024.15.395

learning approaches have been used to predict protein-ligand binding affinities and refine docking poses, offering improved predictive accuracy. Algorithms like AutoDock Vina and DOCK 6 have introduced enhancements in scoring functions and conformational sampling, making docking simulations faster and more accurate. In addition, hybrid docking methods that combine both rigid and flexible docking principles have been developed to balance computational efficiency with accuracy [5].

Conclusion

Protein structure prediction and protein-ligand docking are cornerstone techniques in computational biology, with broad implications for drug discovery, disease modeling and biotechnology. While homology modeling and threading offer robust solutions when homologous templates are available, ab initio methods remain the go-to approach for novel proteins without structural templates. The integration of machine learning into docking algorithms has led to significant improvements in both speed and accuracy and hybrid approaches that combine rigid and flexible docking are becoming more popular for tackling complex systems.

Despite significant advancements in computational techniques, challenges such as protein flexibility, disordered regions and the accuracy of scoring functions still limit the effectiveness of current methods. Nonetheless, ongoing improvements in computational power, algorithm development and data availability promise to further enhance the capabilities of protein structure prediction and protein-ligand docking, ultimately bringing us closer to achieving more accurate and efficient simulations for biomedical and pharmaceutical applications. As the field continues to evolve, the combination of experimental data, enhanced computational models and interdisciplinary collaboration between biologists, chemists and computer scientists will likely pave the way for transformative advancements in our understanding of protein structures and drug interactions.

Acknowledgment

None.

Conflict of Interest

None.

References

1. Xu, Yufei, Feizhen Wu, Li Tan and Lingchun Kong, et al. "Genome-wide regulation of 5hmC, 5mC and gene expression by Tet1 hydroxylase in mouse embryonic stem cells." *Mol cell* 42 (2011): 451-464.
2. Alinejad-Rokny, Hamid, Esmail Sadroddiny and Vinod Scaria. "Machine learning and data mining techniques for medical complex data analysis." *Neurocomputing* 276 (2018).
3. Cai, Chen-Leng, Xingqun Liang, Yunqing Shi and Po-Hsien Chu, et al. "Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart." *Dev cell* 5 (2003): 877-889.
4. Swiercz, Aleksandra, Edmund K. Burke, Mateusz Cichenski and Grzegorz Pawlak, et al. "Unified encoding for hyper-heuristics with application to bioinformatics." *Cent Eur J Oper Res* 22 (2014): 567-589.
5. Allis, C. David and Thomas Jenuwein. "The molecular hallmarks of epigenetic control." *Nat Rev Genet* 17 (2016): 487-500.

How to cite this article: Valentino, Giovanni. "A Comprehensive Review of Algorithms for Protein Structure Prediction and Docking." *Global J Technol Optim* 15 (2024): 395.