Open Horizon Scientific Review

Vol. 03 No.01 2025

Computational Approaches to Predicting Protein Structure and Molecular Interactions

Ayesha Fatima

Department of Plant Sciences, Quaid-i-Azam University, Islamabad, Pakistan

ABSTRACT

The three-dimensional structure of a protein determines its function, yet experimental structure determination remains slow and costly relative to the rapid growth of sequence data. This article reviews computational approaches for predicting protein structure and molecular interactions, beginning with the folding problem and the role of experimental data and databases. Template-based modeling, physics-based and fragment assembly methods, coevolution analysis and modern deep learning systems are compared in terms of principles, strengths and limitations. Extensions of structure prediction to protein-protein complexes, molecular docking of small ligands and molecular dynamics simulation are discussed, together with approaches to validation and confidence estimation. Despite remarkable progress, challenges remain in modeling conformational ensembles, disordered regions, mutation effects, transient interactions and binding affinity. Integration of data-driven prediction with physics-based simulation and experiment is highlighted as a promising direction for understanding protein function and accelerating biomedical discovery.

Key words: *protein structure prediction, deep learning, molecular docking, protein-protein interactions, molecular dynamics*

Introduction:

Proteins carry out most of the functions of living cells, acting as enzymes, receptors, transporters, structural scaffolds and regulators. Their activity arises from the three-dimensional shapes into which amino acid chains fold and from the specific interactions they form with other proteins, nucleic acids and small molecules. Knowing a protein structure is therefore a gateway to understanding mechanisms of health and disease and to designing therapeutics.

Sequencing technologies have produced billions of protein sequences, while experimental methods such as crystallography, nuclear magnetic resonance and cryo-electron microscopy have solved a much smaller number of structures. Computational prediction offers a way to bridge this gap, and the field has been transformed in recent years by the introduction of deep learning systems that can predict structures with accuracy approaching that of experiment (Jumper et al., 2021).

The aim of this article is to provide a coherent overview of computational approaches for predicting protein

structure and molecular interactions. The discussion is organised around nine thematic outlines covering the folding problem, experimental data, homology modeling, fragment and physics-based methods, coevolution, deep learning, complexes, docking and molecular dynamics, followed by an overall summary.

1. The Protein Folding Problem and Computational Prediction

Proteins are linear chains of amino acids that fold into specific three-dimensional structures, and these structures determine their biological functions, from catalysis and signalling to transport and structural support. The classic experiments of Anfinsen showed that the amino acid sequence of a protein contains the information required to specify its native conformation, which implies that structure should in principle be predictable from sequence alone (Anfinsen, 1973).

The protein folding problem has three related aspects: the physical question of how a protein folds, the inverse problem of designing sequences that adopt a chosen fold, and the prediction problem of computing the native structure from sequence. The size of the conformational space makes exhaustive search impossible, and yet real proteins fold on timescales of microseconds to seconds, a paradox that is resolved by the funnel-shaped energy landscape concept in which folding is biased toward the native state. Reviews on the fiftieth anniversary of the problem highlighted both the progress made and the remaining gaps in understanding (Dill and MacCallum, 2012).

Computational prediction is increasingly important because the number of known protein sequences vastly exceeds the number of experimentally solved structures. Reliable models extend structural coverage, support function annotation, enable structure-based drug design and help to interpret the effects of genetic variants. The field is evaluated through blind community experiments in which predictors submit models for proteins whose structures are unknown to them, providing an objective measure of progress (Kryshtafovych et al., 2021).

2. Experimental Structure Determination and Databases

Experimental methods supply the data on which computational methods are trained and tested. X-ray crystallography has produced the majority of known structures, nuclear magnetic resonance spectroscopy provides structures and dynamics in solution for smaller proteins, and cryo-electron microscopy has undergone a resolution revolution that now allows near-atomic structures of large complexes and membrane proteins. These techniques are accurate but slow,

expensive and not always applicable, for example for proteins that resist crystallisation or are intrinsically flexible.

Structural data are archived in public repositories, most importantly the Protein Data Bank, which has grown steadily and now contains hundreds of thousands of entries. Sequence databases are far larger, and families of related sequences collected in multiple sequence alignments carry evolutionary information that has become central to modern prediction. Classification resources organise structures into folds and families and reveal that the number of distinct folds in nature is limited, which underpins template-based modeling.

Data quality matters for computational work. Experimental structures contain errors, missing regions and artefacts resulting from crystal packing or ligand binding, and the Protein Data Bank is biased toward proteins that are easy to study. Training and benchmarking sets must be constructed carefully to avoid information leakage between related proteins, otherwise reported accuracy will be overestimated.

3. Template-Based and Homology Modeling

Template-based modeling exploits the observation that proteins with similar sequences have similar structures, and that structure is more conserved than sequence in evolution. The method identifies a related protein of known structure, aligns the target sequence to it, copies the backbone coordinates for conserved regions, builds loops and side chains for variable regions, and refines the model. An influential implementation derived spatial restraints from the template alignment and then satisfied them to generate a three-dimensional model (Sali and Blundell, 1993).

The accuracy of a comparative model depends mainly on sequence identity between target and template and on the quality of the alignment. Above roughly thirty to forty percent identity, models are typically good enough for many applications, with backbone errors small in the core. At lower identity, alignment errors increase and the model may be unreliable, and in the so-called twilight zone, detection of remote homologues requires sensitive profile and hidden Markov model searches.

Typical errors include misalignment in variable regions, incorrect loop conformations, and inaccurate side-chain packing. Model quality assessment programs estimate the reliability of each model and of regions within it, helping users to decide how far a model can be trusted. Template-based methods remain useful and fast, and they provide a conceptual

foundation for later approaches that implicitly learn the same evolutionary relationships.

4. Physics-Based and Fragment Assembly Methods

When no suitable template exists, ab initio approaches attempt to predict structure from sequence by minimising an energy function. Pure physics-based methods use molecular mechanics force fields and solvent models to search the conformational landscape, but they face two difficulties: the search space is enormous, and the energy function must be accurate enough to distinguish the native structure from near-native decoys. Except for small and fast-folding proteins, these methods have generally been too costly and imprecise for routine prediction.

Fragment assembly offered a practical compromise. Short segments of known structures are selected on the basis of local sequence similarity and are combined through Monte Carlo search guided by a knowledge-based scoring function that captures features such as hydrophobic burial, strand pairing and compactness. The Rosetta software implemented this strategy and achieved notable success in early blind tests for small proteins (Rohl et al., 2004). The same energy functions and sampling methods have been applied to refinement, loop modeling and protein design.

These methods demonstrated that a few principles encoded in local fragments and global compactness can yield approximate folds, but the accuracy was inconsistent for larger proteins and for those with complex topologies. Their value today lies in refinement, design and in situations where physical modeling of conformational change, mutation or binding is needed, often in combination with data-driven components.

5. Coevolution and Contact Prediction

A major conceptual advance came from the use of evolutionary information in multiple sequence alignments. Residues that are in physical contact in the folded structure tend to coevolve, since a mutation at one position is often compensated by a mutation at the contacting position to maintain stability or function. Statistical methods can detect such correlated mutations in a large alignment of homologous sequences, but a naive analysis is confounded by indirect correlations in which two positions appear to covary because each contacts a third.

Global statistical models resolved this problem by disentangling direct from indirect coupling, using approaches related to the inverse Potts model and direct coupling analysis. These predicted contacts could then be used as restraints in folding simulations, and it was shown

that three-dimensional structures of many proteins could be computed from evolutionary sequence variation alone when enough homologous sequences were available (Marks et al., 2011).

The approach is limited by the need for deep alignments, and it performs poorly for families with few known homologues, including many orphan proteins and rapidly evolving proteins. Nevertheless, coevolution established the principle that evolutionary data contain rich structural information and made possible the integration of contact maps with neural networks, setting the stage for the next generation of predictors.

6. Deep Learning Revolution in Structure Prediction

Deep learning changed the field dramatically. Neural networks trained on known structures and sequence alignments were first used to predict inter-residue distances and orientations, and then evolved into end-to-end systems that output atomic coordinates directly. AlphaFold2 combined an attention-based architecture that jointly processes the multiple sequence alignment and pairwise residue representations with iterative refinement of three-dimensional coordinates, and achieved accuracy competitive with experiment for a majority of targets in a blind assessment (Jumper et al., 2021).A parallel system, RoseTTAFold, used a three-track network that passes information among sequence, distance and three-dimensional coordinate representations, and showed that comparable methods could be built in other laboratories and applied also to protein complexes (Baek et al., 2021). Protein language models trained on very large sets of sequences learn evolutionary and structural regularities without explicit alignments, which permits rapid prediction for single sequences and has enabled the generation of hundreds of millions of structural models (Lin et al., 2023).

The conceptual figure below compares the relative accuracy of the main approaches as a function of the similarity of the target to known structures. Template-based methods improve sharply with identity, fragment methods remain modest, and modern deep learning systems maintain high accuracy even at low identity because they draw on large alignments and learned patterns. The figure is illustrative and does not represent measured results, but it conveys the shift in the capabilities of the field.

7. Protein-Protein Interactions and Complex Prediction

Proteins seldom act alone; they form complexes and networks, and predicting how they interact is essential for understanding cellular processes. Computational approaches to

interactions include sequence-based predictors of interface residues, coevolution analysis between pairs of proteins, protein-protein docking that samples relative orientations and scores them, and integrative modeling that combines sparse experimental data from cross-linking, small-angle scattering or cryo-electron microscopy.

Deep learning methods have been extended to complexes. Adaptations of AlphaFold trained specifically for multimers can predict the structure of many protein assemblies and, when run across large sets of candidate pairs, can help to screen for interacting partners (Evans et al., 2022). Prediction is more reliable for stable, well-conserved complexes with deep joint alignments and less reliable for transient interactions, antibody-antigen interfaces and complexes involving disordered regions, for which coevolutionary signals are weak.

Confidence measures, including predicted interface scores and predicted aligned error, are valuable for judging whether a predicted interaction is likely to be genuine. Computational predictions should still be validated experimentally, for example through mutagenesis, binding assays or structural methods, because false positives can arise from non-specific association and from the tendency of models to place chains in plausible contact even when no biological interaction exists.

8. Molecular Docking and Protein-Ligand Interactions

Molecular docking predicts the binding pose and affinity of a small molecule within the binding site of a protein, and is a fundamental tool of structure-based drug discovery. Docking programs combine a search algorithm that explores ligand position, orientation and conformation with a scoring function that ranks poses. Programs such as AutoDock Vina improved the speed and accuracy of docking and made large-scale virtual screening widely accessible (Trott and Olson, 2010).

Docking faces well-known limitations. Proteins are flexible and may change conformation on binding, solvent and water-mediated contacts are difficult to model, and scoring functions are approximate and correlate only moderately with experimental affinities. Strategies to address these issues include ensemble docking against multiple receptor conformations, induced-fit protocols, consensus scoring and post-processing with free energy calculations. Machine learning scoring functions and generative models for pose prediction are being developed and show promise, but they must be tested on truly novel targets to confirm generality.

Predicted structures from deep learning extend docking to targets

without experimental structures. However, the accuracy of side-chain positions and binding-site geometry in such models may not be adequate for precise ligand placement, and they often represent a single conformational state, usually without bound ligands or cofactors. Careful assessment and refinement of binding sites is therefore needed before using predicted models for virtual screening or lead optimisation.

9. Molecular Dynamics, Validation, Limitations, and Future Perspectives

Molecular dynamics simulations solve the equations of motion for all atoms in a system and thereby describe how proteins move, fluctuate and interact with their environment on timescales from femtoseconds to milliseconds. They provide insight into conformational change, ligand binding pathways, allostery and the stability of predicted models (Karplus and McCammon, 2002). Advances in force fields, enhanced sampling techniques and specialised hardware have extended the reach of simulations, and free energy methods can estimate binding affinities with useful accuracy for some systems.

Validation is essential at every stage. Geometric checks of bond lengths, angles, backbone torsions and steric clashes detect obvious errors, while knowledge-based scores assess the plausibility of the fold. Predicted confidence values should be inspected, and where possible models should be compared with experimental data. Blind community assessments continue to play a crucial role in objectively measuring progress and identifying weaknesses (Kryshtafovych et al., 2021).

Important limitations remain. Current methods predict mostly static structures and are less accurate for conformational ensembles, alternative states, intrinsically disordered regions, effects of point mutations, membrane environments, post-translational modifications and interactions with nucleic acids and small molecules. Future directions include prediction of dynamics and ensembles, integration of experimental data, generative design of new proteins and binders, and more reliable estimation of uncertainty. Combining data-driven models with physics-based simulation is likely to provide the most robust understanding of protein function.

Summary

Computational prediction of protein structure has progressed from homology modeling and fragment assembly to coevolution-based contact prediction and, most recently, deep learning systems that deliver near-experimental accuracy for many proteins. These advances rest on the principle that sequence encodes structure

and that evolutionary information in large sequence families captures the constraints on folding. Template-based and physics-based methods retain value for refinement, design and modeling of specific effects.

Prediction has been extended to complexes, ligand docking and dynamics, although accuracy is lower for transient interactions, disordered regions, conformational change and binding affinity. Confidence measures, rigorous validation and experimental testing remain essential. Combining machine learning with physics-based simulation and experimental data is expected to offer the most reliable route toward predicting protein behaviour and supporting drug discovery, enzyme engineering and basic biological research.

References:

- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science, 181, 223-230.
- Baek, M., DiMaio, F., Anishchenko, I., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373, 871-876.
- Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. Science, 338, 1042-1046.
- Evans, R., O'Neill, M., Pritzel, A., et al. (2022). Protein complex prediction with AlphaFold-Multimer. bioRxiv, doi:10.1101/2021.10.04.463034.
- Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583-589.
- Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. Nature Structural Biology, 9, 646-652.
- Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP): Round XIV. Proteins, 89, 1607-1617.
- Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379, 1123-1130.
- Marks, D. S., Colwell, L. J., Sheridan, R., et al. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS ONE, 6, e28766.
- Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein structure prediction using Rosetta. Methods in Enzymology, 383, 66-93.
- Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology, 234, 779-815.
- Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the

speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, 31, 455-461.