JUST AGRICULTURE
multidisciplinary e-Newsletter

Journey to Protein Structure Prediction using A.I.

Gyan Pratap Singh Bhadouriya[1*]
[1]Dept. of Genetics & Plant Breeding, ITM University, Gwalior, (M.P.), india

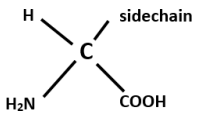
ARTICLE ID: 24

Introduction

The protein structure prediction problem is the question of how a protein's sequence of amino acidsresults in its fully folded 3D structure. protein structures provide a better understanding of the molecular mechanisms of a protein and in doing so offer insights into how the protein works and how its modulationcan help human interest as proteins are one of essential macromolecules working as structural support, transporters, enzymes, gene regulators, receptors etc. in living beings, we know that amino acids are building blocks of proteins, every amino acid contains an amine group, the carboxylic acid and atetrahedral carbon and side chains which dictates the properties of the amino acids, journey of a protein starts from DNA as the flow of information is from DNA to protein, nucleotide sequence of an gene gets transcribed to RNA which after being processed is used by ribosomes to synthesize amino acid chains, ribosomes reads three bases of messenger RNA to add one amino acid to the chain forming a linear heteropolymer referred to as the primary sequence of the protein which determines the structure of the protein, and structure determines the function of protein. All proteins are composed of same 20 amino acids in different composition and sequence. The part what path does primary sequence follows to form 3D structure remains unknown from last 50 years.

3D Structure of Proteins

The primary structure of the proteins is established by peptide bonds and covalent bond's further structuretake place because of non-covalent interaction. Secondary structure is driven by the hydrogen bonding, tertiary structure is by virtue of electrostatic interactions; hydrogen bonds, hydrophobic interactions and van der waals forces or simply we can say interaction between sidechains, sidechains and peptidebackbones, sidechains and water.



JUST AGRICULTURE
multidisciplinary e-Newsletter

All the interactions stated above for formation of secondary and tertiary structure takes place within amino acids of single protein molecule but when these interactions take place between different protein molecules it gives rise to the quaternary structure. Physical chemists, physicists and computational chemists have tried to understand how we can predict the folding structure from primary sequences.

Name	**Interaction**	**Bond (kcal/mol)**
Covalent	Shared electrons	50-100
Ionic	Attraction of opposite charges	Very variable
Hydrogen bond	Interaction of H on N, O, S with electron pair on O, N, S	3-7
Hydrophobic	Interaction of non-polar groups exclusion of water.	1-2
Van der waals	Interaction between weakly polarized bonds	1

Protein Folding

The polypeptide chain of proteins with variable sequence of amino acids makes proteins a heteropolymer that is not constituted of a repeating monomer, the different amino acids with different sidechains attached to them some of which being hydrophobic and others being hydrophilic gives rise to interactions within the polymer causing it fold in a three-dimensional structure and the process whereby you go from the extended primary sequence to the folded structure is called protein folding. In 1944 the physicist Erwin Schrodinger suggested that living systems obeyed all laws of physics and should not be viewed as exceptional but instead reflected the statistical nature of these laws (Whitford, 2013). A major milestonein protein science was the thermodynamic hypothesis of Christian Anfinsen and colleagues he postulated that the native structure of a protein is the thermodynamically stable structure; it depends only on the amino acid sequence and on the conditions of solution, and not on the kinetic folding route (Dill *et al*., 2008). Since then, various aspects of protein folding mechanism have been discovered which appears as a jigsaw. The way Proteins function I would agree with the statement; proteins should be viewed as flexiblemolecules that can take up a whole ensemble of different structural conformations (Van *et al*., 2023). Variussurrounding factors such as temperature, pH, salt concentration etc. affect the structure

JUST AGRICULTURE
multidisciplinary e-Newsletter

of protein making protein folding a random probability distribution or pattern which may be analyzed statistically but may not be predicted precisely. Usually, proteins are present either in their native state (folded state) or unfolded state because of the surrounding factors they continuously fold, unfold and re-fold maintaining adynamic equilibrium.

Thermodynamic Aspects of Folding

In thermodynamics it is assumed that every system will eventually reach a state where there is no net flow of energy or molecules between different parts of the system. To understand it in respect to protein folding *in vivo* (room temperature and neutral pH) proteins are in thermodynamically stable form (native state) since simple environment of proteins are always changing, proteins are found to be in Equilibrium that is individual molecules can still (stochastically) switch between the folded and the unfolded state, but the total number of molecules in each state remains constant. (Van *et al*., 2023). Like any thermodynamic system protein folding also have two free energy minima separated by a barrier, the difference between free energy (the free energy of the folded state is lower that the free energy of the unfolded state.) of both states determine the rate of transmission from one state to another which can be determined by

$$\mathbf{F = H - T\,S}$$

where H is the enthalpy (internal energy in the system)

T is the temperature in Kelvin

S is the entropy (quantification of the amount of conformational freedom of the system).

Protein Folding Problem

In 1969 Cyrus Levinthal, a molecular biologist, estimated that a protein with 100 amino acids, where eachpeptide bond in between two amino acids has two possible torsion angles, and each of these angles can assume three different values. The protein then has $3^{99}\times 2 \approx 2.9 \times 10^{94}$ possible conformations, even ifeach conformation takes one picosecond to go through the confirmation it will take more time then the age of universe (Levinthal's Paradox), and yet, proteins achieve the correct conformation in a fraction of asecond (Al-Janabi, 2022) therefore it can be said that protein folds under some kind of pathways leaving us with questions revolving around its stable state and ways to reach their, *Science* magazine framed the problem this way: "Can we predict how proteins will fold? Out of a near infinitude

JUST AGRICULTURE
multidisciplinary e-Newsletter

of possible ways to fold, a protein picks one in just tens of microseconds. The same task takes 30 years of computer time" (Dill *et al*., 2008).

Conventional Ways of Structure Determination

At the beginning of 2004 over 22 000 protein structures (22348) were deposited in the Protein Data Bank.Over 86 percent of all experimentally derived structures (~19400) were the result of crystallographic studies, with most of the remaining structures solved using nuclear magnetic resonance (NMR) spectroscopy. Slowly a third technique, cryoelectron microscopy (cryo-EM) is gaining ground on the established techniques and is proving particularly suitable for asymmetric macromolecular systems (Whitford, 2013)

Deepmind and Alphafold

Using x-ray crystallography, nuclear magnetic resonance spectroscopy or cryogenic electron microscopy for prediction of protein structure is time consuming and majorly based on trial-and-error approach one ofthe main reasons for which was given by Cyrus Levinthal as stated above, because of which computational procedures are developed for the prediction of 3D protein structures. In 1994 Krzysztof Fidelis (University of California CA, USA) and John Moult (University of Maryland, MD, USA) founded the Critical Assessment of Structure Prediction (CASP) which according to them is an experiment tosolve the half a decade old problem of protein folding, from 1994 CASP takes place in every two years where various teams participates with their computational models for prediction of protein structure, at CASP predictor groups with protein sequences whose structures have been solved but have not yet been made publicly available. To assess the accuracy of predictions made by competitors. The Global Distance Test (GDT) is the main metric used to assess the success of the computational models and ranges from 0 to 100. This can be thought of as the percentage of amino acid residues that are within a certain distance from the correct position, where the experimental structures are used as the 'ground truth' (Al-Janabi, 2022).

Deep Mind has developed an A.I. model that takes us one step closer to predicting the 3D shape of a protein from only its one-dimensional amino acid sequence. without the need for tedious and costly lab analysis. In 2020 Google's Deep Mind introduced Alpha Fold 2.0, which achieved an average GDT of 90 at CASP14, the development of computational methods to predict three-dimensional (3D) protein structures from the protein sequence has proceeded along two complementary paths that focus on either the physical interactions or

JUST AGRICULTURE
multidisciplinary e-Newsletter

the evolutionary history (Jumper et al., 2021) physical interactions as discussed above all the covalent and non-covalent forces in thermodynamic and kinetic stimulations although various interactions are unknown and not been included in alpha fold 2.0 on the other hand evolutionary approach is based on the on bioinformatic analysis of the evolutionary history of proteinsthat is various mutants of similar proteins the makers of alpha fold combined the methods for Alphafold 2.0 it uses the structures available in public domain by world protein banks and data gathered bystructural biologists to make predictions. AlphaFold structures had a median backbone accuracy of 0.96 År.m.s.d.95 (Cα root-mean-square deviation at 95% residue coverage) (95% confidence interval = 0.85–1.16 Å) (Jumper et al., 2021).

Conclusion

No doubt advance in protein structure prediction technology will benefit all the fields of biological sciences and agriculture as well, various aspects of living organisms are either beneficial or non-beneficial to us and these life processes can be altered with the help of knowledge obtained from the technologies like alpha fold. Various diseases like sickle cell anemia are nothing but the structural deformity in proteins, various medicines are designed on the basis of the protein structures, the author believes that there is unseen huge application of this technology in genetics, entomology and plant pathology for development of biotic stress resistant cultivars. also, various knowledge gaps for example the ways in which plant individuals respond to biotic and abiotic stresses can allow us to set more defined breeding programs on the basis of biochemical characters which are strongly correlated with stress resistant qualities of plants.

References

Al-Janabi, Aisha. "Has DeepMind's AlphaFold solved the protein folding problem?" (2022): 73- 76.

Xu, Yongjun, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu et al. "Artificial intelligence: A powerful paradigm for scientific research." *The Innovation* 2, no. 4 (2021): 100179.

Creighton, Thomas E. "Protein folding." *Biochemical journal* 270, no. 1 (1990): 1.

Dill, Ken A., S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. "The protein folding problem." *Annual review of biophysics* 37 (2008): 289.

JUST AGRICULTURE
multidisciplinary e-Newsletter

AlQuraishi, Mohammed. "AlphaFold at CASP13." *Bioinformatics* 35, no. 22 (2019): 4862-4865.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596, no. 7873 (2021): 583-589.

Chan, H. S., & Dill, K. A. (1993). The protein folding problem. *Physics today*, *46*(2), 24-32.

Englander, S. W., & Mayne, L. (2014). The nature of protein folding pathways. *Proceedings ofthe National Academy of Sciences*, *111*(45), 15873-15880.

Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics? *Protein Science*, *28*(4), 678-683.

Whitford, D. (2013). *Proteins: structure and function*. John Wiley & Sons.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444.

van Gils, J. H., van Dijk, E., May, A., Mouhib, H., Bijlard, J., Jacobsen, A., ... & Abeln, S. (2023). Introduction to Protein Folding. *arXiv preprint arXiv:2307.02174*.

van Gils, J. H., Mouhib, H., van Dijk, E., Dijkstra, M., Houtkamp, I., Goetzee, A., ... & Feenstra, K. A. (2023). Thermodynamics of Protein Folding. *arXiv preprint arXiv:2307.02175*.