Artificial Neural Networks Applied to Protein Structure Determination
A Review

Jack Goon
University of California, Davis
UWP 104 - Writing in Science

I. Introduction

Protein structure and folding is foundational to biology because life depends on metabolic pathways composed of protein enzymes, transcription factors, and more. Investigation into protein structure started with with the X-ray crystallography of myoglobin in 1960, where a protein's general structure was experimentally determined to with a 2-Angstrom resolution [1]. Shortly after, Anfinsen proposed the "thermodynamic hypothesis" that proteins fold according to their amino acid sequence to the lowest energy conformation [2]. This hypothesis started a decades-long hunt for a model that can read an amino acid sequence and produce a fully formed three-dimensional structure. Early computer simulations tried to find the lowest energy conformations, but the conformation space for proteins increased exponentially and beyond the capabilities of computers [3] [4]. To circumvent this large conformation space, simplifications are made. Perhaps the most influential is the lattice model. Under the first lattice model, a two-dimensional square grid represents the protein conformation. At each "node" of the grid, an amino acid is present and bonding with the adjacent nodes. While this model reduces the simplified the conformational space, the problem remains exponential [4].

Around this same time, the first computers could begin implementing "self-organizing" information systems [5]. These systems, which later became known as artificial neural networks, are trained with data to adjust the weights of the "neurons." Once trained, new data can be fed through the neurons to produce results with accuracy [3]. B. G. Farley, the researcher who first implemented artificial neural networks, notes that these networks are especially good at solving problems too complex for simple analytical solutions [5]. As we will see, the complex protein folding problem is no exception. This paper will explore what types of neural networks exist to predict protein secondary and tertiary structure from a primary structure.

II. Overview of Artificial Neural Networks

First, I will review the different types of neural networks. Neural networks are composed of artificial neurons. Each neuron has a set of inputs (x), each connected by a weight (w), a threshold function (Φ) and an output (y) as described by McCulloch and Pitts in 1943 [6] [**Figure 1 a**]. These neurons can be grouped into layers in a forward feeding neural network (FFNN), where the outputs of layer k are the inputs to every neuron in layer k+1. The first layer is an input layer, followed by numerous hidden layers, and lastly an output layer to be interpreted as the result of the neural network [**Figure 1 b**]. The values of these layers are updated with the backpropagation algorithm, which takes into account how well the model performed [7]. While these networks are quite simple, they can be combined in ways that provide flexibility for researchers [3]. This will become clear once we analyze protein folding-specific networks. Convolutional neural networks (CNN) add additional layers to the beginning to extract relevant features. This is a process that previously required expert intervention [9]. Convolutional neural networks are useful in image analysis, where certain pixels make up a “feature” in an image. For the purpose of this paper, we can consider convolutional neural networks a subset of forward feeding neural networks. Lastly, Recurrent neural networks (RNNs) are best for sequential inputs. They have a memory buffer that remembers the previous output values while processing the current input. This memory buffer serves as an additional input alongside the actual input value. This allows the network to consider previous samples, which could be important for ordered data [10].

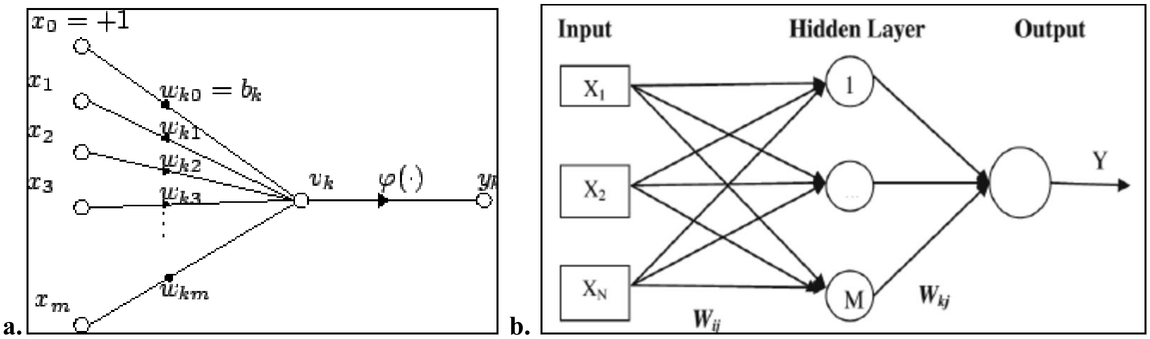


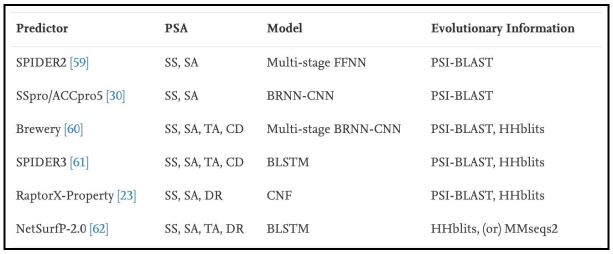
Figure 1
a. Diagram of the k-th artificial neuron in a neural network layer, where v is the neural node. The output of the node can be described by $y[k] = \varphi(\sum_{i=0}^{m} w[k,i] * x[k])$ where phi is a threshold function.
Diagram Source: “Artificial Neuron,” Wikipedia.com.
b. A forward-feeding neural network [8]

III. Predicting 1D and 2D Protein Structure Annotations

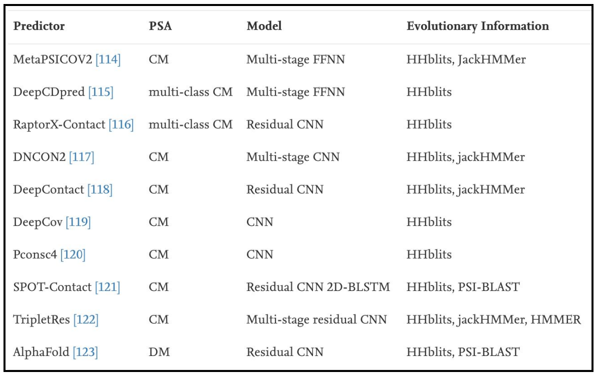
Ideally, a full three-dimensional structure could be directly derived from an amino acid sequence with a single neural network. This will be discussed in section **IV.** Unfortunately, these systems are not fully developed. The current state-of-the-art solution is to combine specialized neural networks in a pipeline [3]. These pipelines might integrate solutions from two general approaches to protein structure representation. First we will describe one-dimensional protein structure annotations. These annotations are considered one-dimensional because they can be mapped onto amino acid sequence. Secondary structure (alpha helices, beta sheets, and several other specific structures), solvent accessibility (the contact between each residue and the surrounding solvent), torsional angles (describing the protein backbone conformation), and more can all be described one-dimensionally [11]. Next, two-dimensional protein structure annotations include contact maps, which describe the contacts between amino acids for each amino acid pair in the protein. This is two-dimensional because it requires a matrix representation rather than an array representation [11]. Now, we will go into detail on these two different approaches.

III-A. One Dimensional Protein Structure Annotations

First, we will look at the most accurate and innovative attempts at predicting one-dimensional protein structure annotations. A list of deep learners designed for one-dimensional protein structure annotations can be found in **Figure 2 a.** NetSurfP-2.0 follows a fairly straightforward architecture with a few important subtleties. It is a convolutional neural network with two bi-directional layers (called Long Short-Term Memory or LSTM) following the convolutional layers. The authors claim that these bi-directional layers, more so than the convolutional layers, contribute to the network success [12]. The network has a Q3 (helix, strand, and coil prediction) accuracy of 85% [12]. Bi-directional recurrent neural networks (BRNNs) were first invented by Schuster and Paliwal, and are similar to recurrent neural networks described earlier. They function by taking into account both the previous output values and the future output values while determining the current output. Thus, the network gives temporal context to each adjustment [13]. In biological terms, this means that there is more context for each amino acid in the protein [12]. Rather than giving context in a temporal direction, as originally intended, the BRNN acts in a spacial direction for protein structure neural networks. Another program, SPIDER3, uses Long Short-Term Memory and sees a 3% increase in accuracy for a total of 84% accuracy [14]. The theoretical limit for Q3 accuracy is 88 to 90% [15].



Predictor	PSA	Model	Evolutionary Information
SPIDER2 [59]	SS, SA	Multi-stage FFNN	PSI-BLAST
SSpro/ACCpro5 [30]	SS, SA	BRNN-CNN	PSI-BLAST
Brewery [60]	SS, SA, TA, CD	Multi-stage BRNN-CNN	PSI-BLAST, HHblits
SPIDER3 [61]	SS, SA, TA, CD	BLSTM	PSI-BLAST, HHblits
RaptorX-Property [23]	SS, SA, DR	CNF	PSI-BLAST, HHblits
NetSurfP-2.0 [62]	SS, SA, TA, DR	BLSTM	HHblits, (or) MMseqs2



Predictor	PSA	Model	Evolutionary Information
MetaPSICOV2 [114]	CM	Multi-stage FFNN	HHblits, JackHMMer
DeepCDpred [115]	multi-class CM	Multi-stage FFNN	HHblits
RaptorX-Contact [116]	multi-class CM	Residual CNN	HHblits
DNCON2 [117]	CM	Multi-stage CNN	HHblits, jackHMMer
DeepContact [118]	CM	Residual CNN	HHblits, jackHMMer
DeepCov [119]	CM	CNN	HHblits
Pconsc4 [120]	CM	CNN	HHblits
SPOT-Contact [121]	CM	Residual CNN 2D-BLSTM	HHblits, PSI-BLAST
TripletRes [122]	CM	Multi-stage residual CNN	HHblits, jackHMMer, HMMER
AlphaFold [123]	DM	Residual CNN	HHblits, PSI-BLAST

Figure 2

a. 1D PSA deep learners [3]

b. 2D PSA deep learners [3]

III-B. Two Dimensional Protein Structure Annotations

While one-dimensional annotations provide useful information on a protein, they only supply information for each amino acid individually and say nothing about broader protein structure. Two-dimensional contact maps, however, provide useful information about the locations and interactions between amino acids that could be on opposite sides of the polypeptide chain. A contact map is a matrix with “1” values if amino acids are close enough to each other (a typical threshold value is 8 Angstroms) and a “0” value if they are not. This binary representation is not a comprehensive two-dimensional representation of a three-dimensional structure [16]. Though it is not directly three-dimensional, there exists programs like GDFuzz3D to convert these contact maps into three-dimensional structures. GDFuzz3D is a two-part program that converts the contact map into a map with euclidean distances between each amino acid. It then continues to convert that distance map into a three-dimensional structure [16].

Since researchers use two-dimensional contact maps to create three-dimensional protein structure representations, these contact maps are of extreme interest. **Figure 2 b** lists the major deep learning attempts to create contact maps. Early attempts used bidirectional recurrent neural networks, which were discussed in section **III-A** [17]. Predictions by Pollastri and Baldi have 60.5% accuracy for 8 angstrom cut offs, which is not ideal. However, there are often evolutionarily-close proteins (homologous proteins) that can help guide predictions. These are called templates. One attempt uses templates with 20% identity (median) and sees 1.2 angstrom improvements over weighted average models [18]. RaptorXContact, and many other deep learning programs that followed it, use convolutional neural networks to predict two dimensional contact maps [19]. RaptorXContact uses 60 convolutional layers to achieve its goal [19].

IV. End-to-End Approaches

So far, we have described pathways from amino acids to one-dimensional or two-dimensional structural annotations. Now I will discuss how neural networks help reach the end goal of a three-dimensional structure. I have already described how contact maps can be converted into these three-dimensional structures [16]. However, there are some more comprehensive pathways available, including some that are not fully developed but show more promise.

The first approach is to combine multiple solutions together in a pipeline. AlphaFold, a system entered into CASP13, combines three neural networks that do not use any templates, unlike some systems discussed in section **III**. One creates a euclidean distance matrix (a two-dimensional representation), one estimates structure accuracy, and the last generates the protein structures [20]. The program uses simulated annealing to combine the results. Simulating annealing works by inserting a fragment into a structure and accepting or rejecting that fragment based on certain criteria [20] [21].

Other fields, like speech recognition and computer vision, are moving towards an end-to-end differentiable approaches that are not just a combination of smaller solutions [23]. Protein structure prediction is no exception. Recurrent geometric networks (RGNs) aim to start with amino acid structures and end with a fully formed three-dimensional structure [22]. This is done by feeding amino acids into the structure one at a time and updating torsional angles and geometry [22]. While no secondary structure is fed into the neural network, the system was found to create its own secondary structure representations [22]. See **Figure 3** for a diagram of the recurrent geometric network architecture. These networks significantly speed up structure prediction but take longer to train [22]. Hopefully, as the field progresses, more examples of recurrent geometric networks will improve on the first model provided by AlQuraishi.

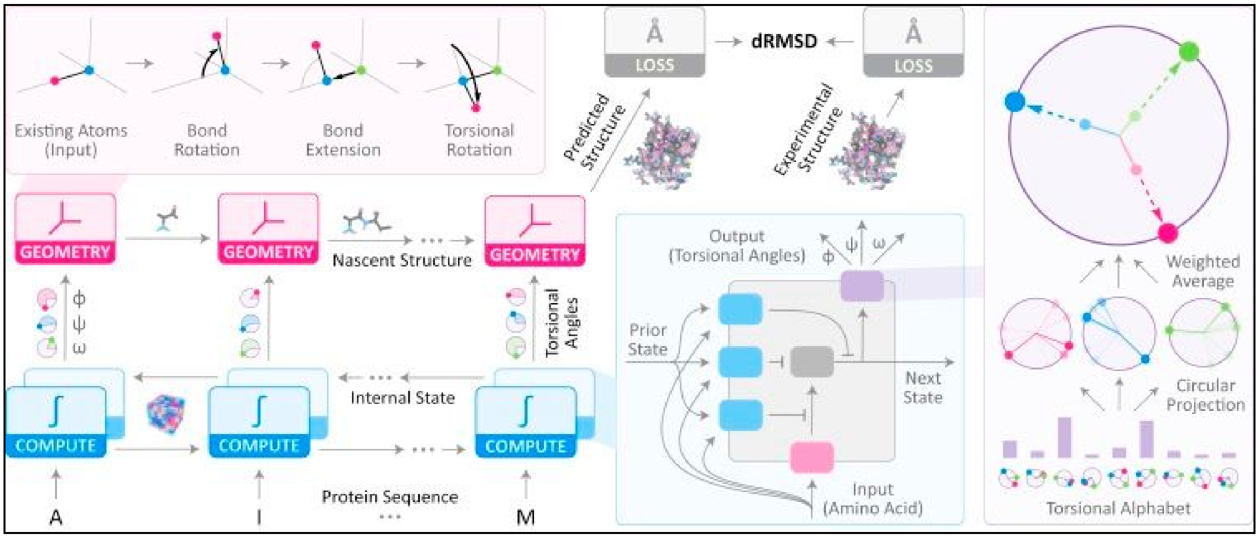


Figure 3
A recurrent geometric network that iteratively computes torsional angles, updates geometry, and compares to experimental structures to provide loss values that will be used to train network parameters [22]

V. Conclusion

Neural networks offer a promising solution to one of biology's hardest problems. While they can only achieve certain levels of accuracy, deep learning systems have replaced other methods as the state-of-the-art solution [3]. One dimensional protein structure annotation networks are approaching their theoretical limits [15]. Two dimensional proteins structure annotations are still being improved and can accurately inform three dimensional structures. Lastly, pioneers are currently creating end-to-end differential systems like the recurrent geometric network, which parallel the transitions made in speech recognition and computer vision [22]. While there is plenty of research and innovation to be had, protein folding neural networks have come a long way since the invention of the neural network in the early 1960s. In the future, these networks will reduce our reliance on X-ray crystallography and enhance our ability to create novel proteins for drugs, vaccines, and industry.

VI. Bibliography

1. Kendrew, John C., et al. "Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å. resolution." *Nature* 185.4711 (1960): 422-427.
2. Anfinsen, Christian B. "Principles that govern the folding of protein chains." *Science* 181.4096 (1973): 223-230.
3. Le, Quan, Mirko Torrisi, and Gianluca Pollastri. "Deep learning methods in protein structure prediction." *Computational and Structural Biotechnology Journal* (2020).
4. Lau, Kit Fun, and Ken A. Dill. "A lattice statistical mechanics model of the conformational and sequence spaces of proteins." *Macromolecules* 22.10 (1989): 3986-3997.
5. Farley, B. W. A. C., and W. Clark. "Simulation of self-organizing systems by digital computer." *Transactions of the IRE Professional Group on Information Theory* 4.4 (1954): 76-84.
6. McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics* 5.4 (1943): 115-133.
7. Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *nature* 323.6088 (1986): 533-536.
8. Yılmaz, Ali, Cigdem Aci, and Kadir Aydin. "MFFNN and GRNN Models for Prediction of Energy Equivalent Speed Values of Involvements in Traffic Accidents/Trafik Kazalarında tutulumunun Enerji Eşdeğer Hız Değerleri Tahmininde MFFNN ve GRNN Modelleri." *International Journal of Automotive Engineering and Technologies* 4.2 (2015): 102-109.

9. LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.
10. Elman, Jeffrey L. "Finding structure in time." *Cognitive science* 14.2 (1990): 179-211.
11. Torrisi M., Pollastri G. (2019) Protein Structure Annotations. In: Shaik N., Hakeem K., Banaganapalli B., Elango R. (eds) Essentials of Bioinformatics, Volume I. Springer, Cham
12. Klausen, Michael Schantz, et al. "NetSurfP‐2.0: Improved prediction of protein structural features by integrated deep learning." *Proteins: Structure, Function, and Bioinformatics* 87.6 (2019): 520-527.
13. Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11 (1997): 2673-2681.
14. Heffernan, Rhys, et al. "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility." *Bioinformatics* 33.18 (2017): 2842-2849.
15. Yang, Yuedong, et al. "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?." *Briefings in bioinformatics* 19.3 (2018): 482-494.
16. Pietal, Michal J., Janusz M. Bujnicki, and Lukasz P. Kozlowski. "GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function." *Bioinformatics* 31.21 (2015): 3499-3505.
17. Pollastri, Gianluca, and Pierre Baldi. "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners." *Bioinformatics* 18.suppl_1 (2002): S62-S70.

18. Kukic, Predrag, et al. "Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks." *BMC bioinformatics* 15.1 (2014): 6.
19. Wang, Sheng, et al. "Accurate de novo prediction of protein contact map by ultra-deep learning model." *PLoS computational biology* 13.1 (2017): e1005324.
20. Senior, Andrew W., et al. "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)." *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019): 1141-1148.
21. Kirkpatrick, Scott, C. Daniel Gelatt, and Mario P. Vecchi. "Optimization by simulated annealing." *science* 220.4598 (1983): 671-680.
22. AlQuraishi, Mohammed. "End-to-end differentiable learning of protein structure." *Cell systems* 8.4 (2019): 292-301.
23. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.