Lecture Title and Date

Computational Modeling of Protein Structure, 4/16

Objectives of the Lecture

By the end of the lecture, students should be able to:

1. Explain the nature of the **protein folding problem**
2. Understand **wet-lab techniques** for protein structure determination (X-ray crystallography, Cryo-electron microscopy)
3. Understand **dry-lab approaches** to protein structure prediction (AlphaFold)
4. Identify the **advantages and disadvantages** of using wet-lab vs dry-lab approaches

Key Concepts and Definitions

Protein folding problem – elucidating protein structure from amino acid sequence

Critical Assessment of Structure Prediction (CASP) competition – a recurring competition among researchers for development of the best structure prediction model

X-ray crystallography – a method for structure elucidation that analyzes patterns from X-ray diffraction of protein crystals

Nuclear magnetic resonance (NMR) spectroscopy – a method for structure elucidation that measures shifts caused by the local environment to reconstruct structure

Cryo-electron microscopy (Cryo-EM) – a method for structure elucidation that uses transmission electron microscopy to capture views of a protein from different angles

AlphaFold – a series of models from Google DeepMind for computational structure prediction

Multiple sequence alignment (MSA) – identifies similarities/differences between 3+ sequences

Convolutional neural network (CNN) – a neural network architecture that uses convolutional filters to learn features from an image

Transformer – a neural network architecture popularized in natural language processing that models sequential data using an encoder-decoder

Self-distillation – a transfer learning approach that allows an AI model to “teach” itself, i.e., identify the most important aspects of a large model to make a smaller model with similar performance

Protein Data Bank (PDB) – a repository of elucidated protein structures

Main Content/Topics

Protein folding problem

The problem of predicting protein folding is complicated by two factors: (1) a given sequence could have multiple low-energy folded representations, and (2) protein structure can depend on environmental conditions. Proteins can take on different quaternary structures, from single monomers to homomultimers to complexes. In the years since the first elucidation of protein structure (myoglobin), progress has been made to develop 3 main wet-lab techniques for structure determination. X-ray crystallography applies X-rays to a crystallized structure in order to study the diffraction patterns, NMR uses the electromagnetic shifts to determine constraints/reconstruct protein structure, and Cryo-EM captures different perspectives of a protein to recover the true structure. Given these major advancements, the number of structures available in the Protein Data Bank (PDB) has exploded to over 200,000.

Prediction of protein structure

Given the cost and difficulty involved in wet-lab protein structure determination, prediction of protein structure is a key challenge for computational scientists. In a large-scale prediction competition called CASP, performance has improved drastically in the past decade, culminating in the release of AlphaFold2 in 2020. AlphaFold2 is a deep learning model that uses a variety of components. In structure prediction models like AlphaFold2, the input is the result from multiple sequence alignment or a structure template, the model is a neural network architecture, and the output can be constraints or coordinates. Multiple sequence alignment is valuable because it can suggest whether residues are in the core or on the surface (core residues are likely conserved, while surface residues are not). Three possible input formats are shown below.

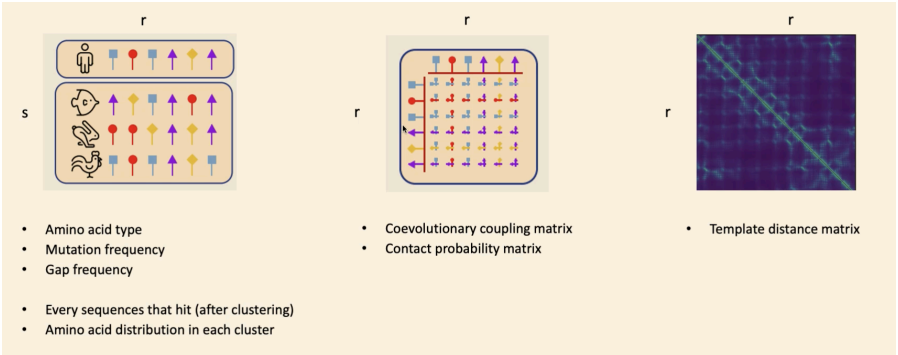

Fig. 1. A r by s matrix showing multiple sequence alignment across species (left), a r by r square matrix consisting of contact probabilities (middle), and an r by r square distance matrix inferred from the probabilities using actual structure (right).

Development of AlphaFold

The original version of AlphaFold, AlphaFold1, cast the structural prediction as an "image recognition problem." Based on the an r by r input matrix, they applied a convolutional neural network (CNN) for prediction of the contact probability between pairwise residues. Based on these probabilities, an energy function was constructed and minimized in order to simulate the process of folding. AlphaFold2, which optimized AlphaFold1, demonstrated strong agreement between predictions and ground truth for most small proteins and best-in-class performance in the CASP competition. The AlphaFold2 architecture (shown in Figure 2) frames the problem as an end-to-end "language processing problem" rather than an "image recognition problem." It operates on the multiple sequence alignment and pairwise MSA data, passing these data into an transformer-based processing block and then into a structure block that infers a predicted structure. Further studies of AlphaFold2 showed self-distillation improves results and that the inclusion of a template does not improve performance, demonstrating model robustness. AlphaFold3 (shown in Figure 3) used a diffusion module to allow for improved resolution of complexes like protein-ligand structures.

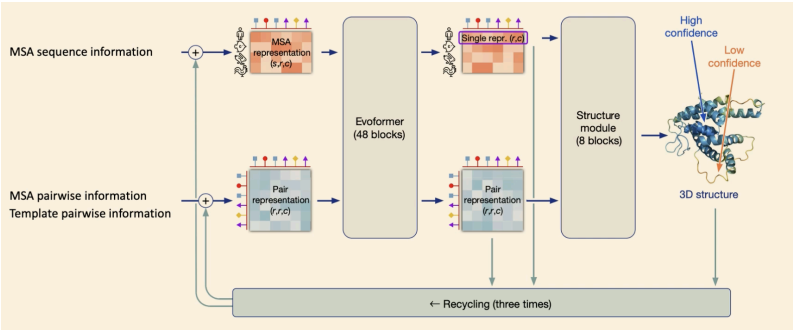

Fig. 2. AlphaFold2 model architecture, featuring a transformer-based Evoformer for processing and structure model for inference.

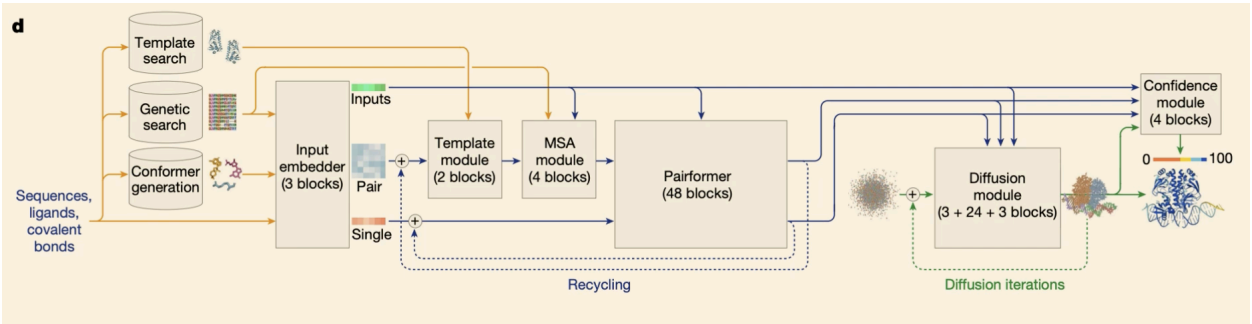

Fig. 3. AlphaFold3 model architecture, featuring various processing steps and a diffusion module to denoise data and infer a structural prediction.

Limitations of structure prediction

Depth of the multiple sequence alignment is a key limitation of AlphaFold models: poor depth results in inaccurate predictions. Additionally, AlphaFold cannot model the impact of mutations on structure or generate a sequence that will give rise to a certain protein ("protein design problem"). The protein design problem represents a major challenge in computational science because it requires out-of-sample generalization. Even the newest models like AlphaFold3 continue to struggle with the modeling of protein-RNA, protein-DNA, protein-protein, and protein-ligand interactions, despite promising performance in monomer prediction. These results are consistent with the difficulty of interaction prediction. Further, these models can be sensitive to the random seeds used. Furthermore, there may be gaps between *in vitro* ground truths determined by X-ray crystallography or Cryo-EM and *in vivo* ground truths that represent the true protein structure in physiological conditions.

Discussion/Comments

This lecture provided a strong overview of developments in protein structure prediction motivated by the expense and difficulty in obtaining structures via traditional methods (X-ray crystallography and Cryo-EM). AlphaFold1 cast structure prediction as an "image recognition problem," and AlphaFold2 improved performance for monomers by framing it as a "language processing problem" instead. AlphaFold3 further improved performance–especially for complex elucidation–using a diffusion model.

Yet there are two main limitations to the state-of-the-art computational approaches. One is that what is considered to be ground truth in training these models is derived from *in vitro*, rather than *in vivo*, contexts. The ground truths arising from X-ray crystallography and Cryo-EM may not reflect true physiological structure due to steric crowding and non-steric effects. Overcoming this obstacle likely requires novel approaches to capturing protein structure *in vivo*. Another key limitation is that current methods of evaluating structural predictions are somewhat arbitrary. More work is needed to develop a good scoring function that can measure the "goodness of fit" for predicted structures.

Despite these limitations, dry-lab approaches have drastically sped up biophysical research by providing a first draft of protein structure which can be confirmed in wet-lab experiments. AlphaFold3's improved ability to model complexes has major implications for the design of new therapeutics and for understanding the impact of the proteome on human disease processes.