CSE474 - SIMULATION AND MODELING

MID TERM PAPER REVIEW

SAMIHA AFAF NEHA
ID 20101266

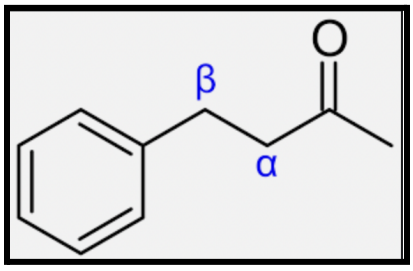
GENERATIVE MODELING FOR PROTEIN STRUCTURES

Namrata Anand PO-Ssu Huang

Background

Proteins are complex molecules that are vital to the structural and functional mechanisms of our body. These are made up of long chains of building blocks known as amino acids. Twenty types of acids interact with each other, forming peptide bonds and eventually constituting what is known as the protein backbone with characteristic side chains. These side chains, along with the amino acid sequence, interact with each other, constructing the 3D secondary structure (e.g., coil, helix, beta pleated sheets) of a protein. This is known as protein folding, a domain that has been the subject of extensive research and modeling to decipher the folding pathways and mechanisms leading to the discovery of the native protein structure. This deep understanding of biology has led to tremendous mileage in the discovery of new therapies, enzymes, small-molecules and bio-binders.

The paper "Generative Modeling for Protein Design" primarily deals with this classic problem of protein folding or protein structure prediction through modeling protein sequences leading to the evaluation of the secondary structure. Unlike the previous approaches, where one had to know the amino acid sequence for structure generation, this paper proposes a generative model based on a new elementary trait of the protein chain - the pairwise distances of α - carbons along a peptide backbone.



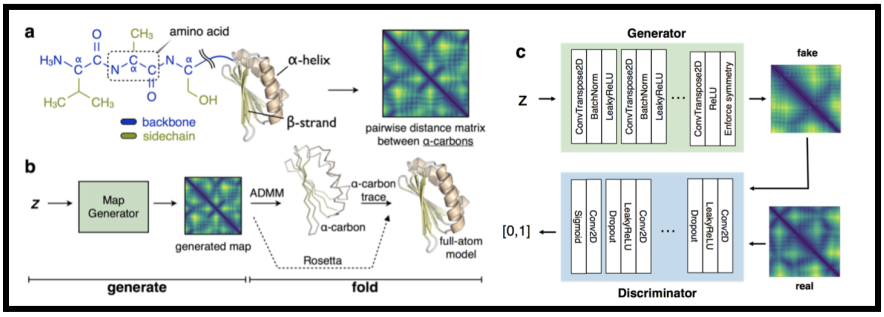
In a protein structure, alpha carbon is the carbon attached to the carbonyl functional group (shown above) and is the central point in the backbone of the amino acid. This paper, anchoring on this trait, uses deep learning and optimization algorithms to extrapolate the 3D protein structure based on the pairwise distance metric of the α - carbons in a protein chain. This novel approach has been experimentally proved to be more accurate, fast and most importantly invariant to the translation and rotational symmetries of the macromolecule, making it simpler.

Abstract Model

Although previously proteins were known to fold based on the amino acid sequence, in recent times, it has become more apparent that protein-protein interfaces also conform largely based on distinct structural motifs. This led authors to question whether there existed such a motif trait that would enable them to fabricate a sequence-agnostic protein structure generation model. Consequently,they came up with the pairwise α - carbon distance approach where primarily the protein structures are represented as pairwise distances (in angstroms) between the α - carbons. This data is then processed sequentially using two different models for evaluating structure folding and sequence inpainting. These are,

1. The deep convolutional adversarial model or DCGANs (classifies as a deep learning generative model)

2. Alternate Direction Multipliers or ADMM (convex optimization problem solving algorithm)



The DCGAN is used to generate the pairwise distance maps from the input data. The generator part of the GAN takes in a random vector z and outputs a fake distance map to delude the discriminator, which then deduces whether the inputs are genuine or fake. After the pairwise distance matrix generation, ADDM is used to "fold" the 2D pairwise distance map into 3D cartesian coordinates of the α - carbons, making a 'trace script' i.e., the alpha carbon coordinate position, which is used to trace a plausible protein backbone through the α - carbon coordinates.

Mathematical Model

For the GAN model,given a $\mathbf{z} \sim \mathcal{N}(0, I)$ vector and data x, the generator G and the discriminator D aim to maximize the following objectives to generate new data maps concurrent with the input data maps given.

$$\max_{D} \ \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-(D(G(\mathbf{z}))))]$$
$$\max_{G} \ \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(D(G(z)))] \tag{1}$$

For the convex formulation of the ADMM algorithm, the gram matrix that will eventually give us the cartesian coordinates of the alpha carbon has been computed using the semidefinite program SDP. A slack term η (found using several iterations of gradient descent) is also introduced for each distance where the L1 norm is penalized by a weight λ, in order to make the model robust to corruptions in distance measurements. Finally the ADMM formulation is modified for optimization which is decomposed over iterative updates over the variables G, Z, U as shown below,

$$\min_{G,\eta} \lambda \left\|\eta\right\|_1 + \frac{1}{2}\sum_{i=1,j=1}^{m}(g_{ii}+g_{jj}-2g_{ij}+\eta_{ij}-d_{ij}^2)^2$$
$$\text{subject to} \quad G \in \mathcal{S}_+^n \tag{2}$$

$$\min_{G,Z,\eta} \ \lambda \left\|\eta\right\|_1 + \frac{1}{2}\left(\sum_{i=1,j=1}^{m}(g_{ii}+g_{jj}-2g_{ij}+\eta_{ij}-d_{ij}^2)^2\right) + \mathbb{1}\{Z \in \mathcal{S}_+^m\} \tag{3}$$
$$\text{subject to } G - Z = 0$$

$$G_{k+1}, \eta_{k+1} = \underset{G,\eta}{\text{argmin}} \ [\ \lambda \left\|\eta\right\|_1 + \quad \frac{1}{2}\sum_{i=1,j=1}^{m}(g_{ii}+g_{jj}-2g_{ij}+\eta_{ij}-d_{ij}^2)^2 + \frac{\rho}{2}\left\|G - Z_k + U_k\right\|_2^2\ ]$$
$$Z_{k+1} = \Pi_{\mathcal{S}_+^n}(G_{k+1}+U_k)$$
$$U_{k+1} = U_k + G_{k+1} - Z_{k+1} \tag{4}$$

Method

The data used in the research was collected from the Protein Bank, which is an online repository of experimentally found protein structures. To eliminate the need for transformational encodings, this type of protein structure representation was used instead of the full-atom, high resolution structure available. Next, the 3D structure was encoded into 2D pairwise distance maps between the α - carbons. This presentation does not contain the torsion angle or side chain information of the peptide chain but has enough information for a complete structure recovery as the maps preserve the order of the peptide chain from the N- to C- terminus which is fixed in peptide chains.

To reduce structural homology in GAN maps, the training and testing datasets were divided using the SCOP (Structural Classification of Proteins) fold type. This ensured that the GAN was indeed generating new maps.

The dataset was created by extracting non-overlapping segments of length 16, 64, and 128 residue from chain ‘A’, which was used to train the GAN. The data was split into training and testing sets for corruption recovery experiments and not as part of the GAN model itself.

For folding generated maps, two methods were tested. First, the Rosettas’ optimization toolkit, to find backbone sequence via fragment sampling. But this method was not scalable in folding and estimating longer structures due to fragment and rotamer (a conformational isomer) sampling rates. Thus, the faster ADMM approach was used for computation. PyTorch was used for the coding implementation of the model. For inferring the 3D structures from the ADMM output, the Rosetta modeling suit was used.

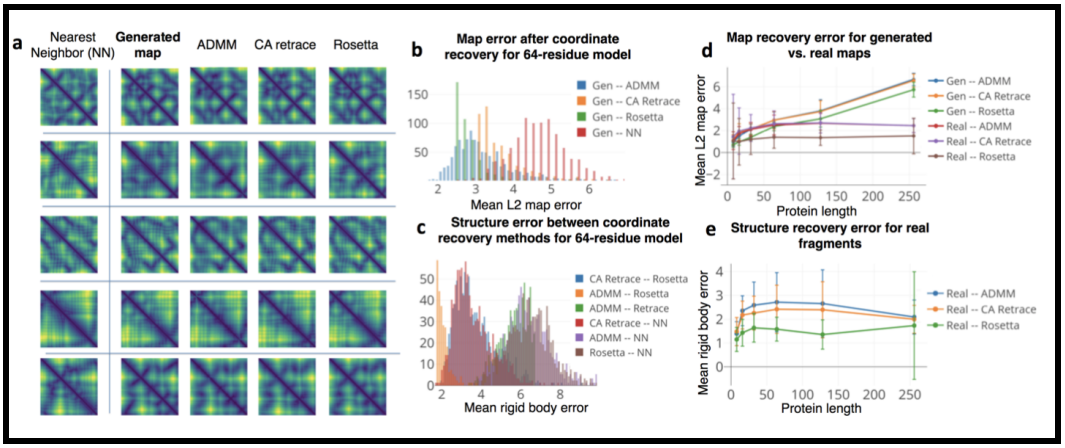
Baseline Models For The Folding Task

To compare the efficiency and novelty of the proposed model, few baseline models were taken into consideration as listed below,

- Rosetta based fragment-sampling model
- TorusDBN and FB5-HMM - uses Hidden Markov Models to generate local backbone torsion angle and α - carbon coordinate placement.
- Use of SDP to infer protein structure using Magnetic Resonance Imaging.
- Residue-coevolution based structure prediction - this requires knowledge of the amino acid chain.

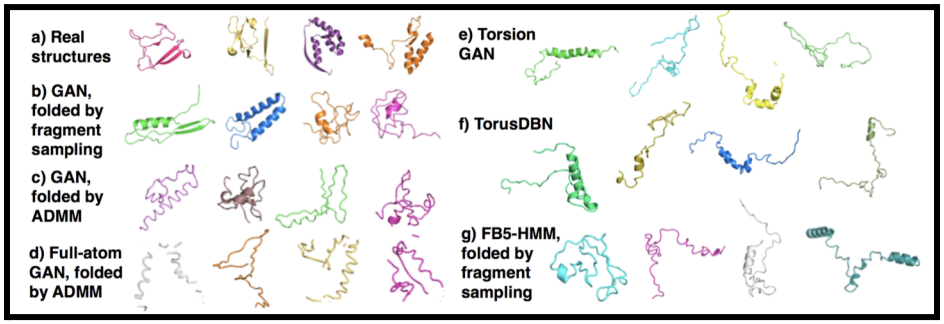
Results and Output Validation

After generating 16, 64, and 128 residue maps using GAN and then folding them using ADMM and Rosetta protocols, the model is compared to baseline models such as the multi-scale torsion angle GAN, 3D GAN, full-atom GAN, TorusDBN, and HB5HMM. A plot of the mean L_2 map error and the rigid body error shows that the α - carbon approach performs better (as shown below) than the baseline models stated earlier.



For the folding of the structure, it was observed that the generator in DCGAN was able to reconstruct meaningful secondary structural instances, e.g., the beta pleated sheets, alpha helix, validating the proposed model. Folding using Rosetta model was slower than the α - carbon approach but could correct unusual error in the local structure of the protein.The other baselines also underperform relative to the pair-wise

α - carbon distance method by looping onto themselves or giving rise to unrealistic protein 3D structure. The folding of each model is shown below for comparison.

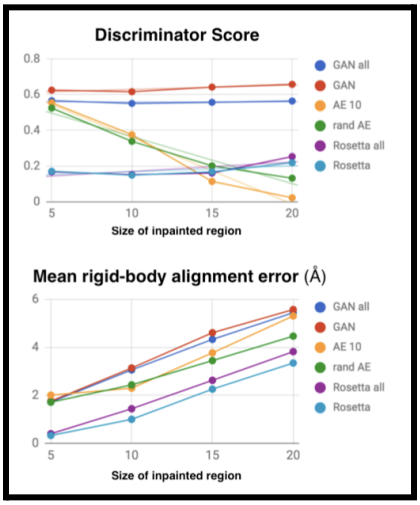


Inpainting and Inpainting Results

Another problem addressed by the paper is protein inpainting, where the model correctly tries to infer missing portions of a protein structure, especially in loop modeling. For experiment purposes, a few subsets of the pairwise distance are eliminated, and the model is used to fill in the missing distances based on the context provided by the uncorrupted sections.

Just like structure fold estimation, for inpainting, some baseline models such as the 10-residue supervised encoder, RosettaRemodel, and Random Corruption supervised autoencoder were taken into consideration. As the paper suggests, there is no canonical evaluation metric to determine the inpainting accuracy, but

one probable approach could be to compare the discriminator score and the rigid-body alignment error against the native structure.



As shown in the figures above, the labels "GAN all" (the result for 10 solutions per structure) and "GAN"' (the top solution per structure) show the highest discriminator score but deviates most in the rigid body alignment error plot. On the other hand, although autoencoders produce faster inpainting results, they produce unrealistic solutions when inpainting over longer regions of a peptide chain .The GAN has the advantage over autoencoder in that it can handle arbitrary loop closures and sample multiple solutions for a single inpainting task.

Final Review

The already existing Rosetta fragment sampling method with high energy function gave accurate protein folding deductions, but the slowness of the fragment sampling stage and the number of sample steps needed to reach practical solutions were this method's key flaws. In this regard, the paper offers a quick technique for loop modeling or fold deduction, using a generative model that considers the overall protein structure without having to know the actual amino acid sequence.

In conclusion, the paper has devised a state-of-the art model incorporating the α - alpha carbon distance feature of a peptide backbone to solve the protein folding and inpainting problem. By generating pairwise distance maps using DCGAN and then building the α - carbon trace using ADMM algorithm, the trace script is then fed into the Rosetta modeling kit to finally deduce new conformations. Regardless of the better solutions this model poses relative to the baselines, the paper also states a few scopes for extending the work for accuracy, such as input data conditioning for improved structure generation and GAN performance enhancement for semi-supervised prediction tasks.