Reconstructing protein structure from contact map

CHEN, Tianhao, Dept. of Mathematics, HKUST, tchenbb@connect.ust.hk

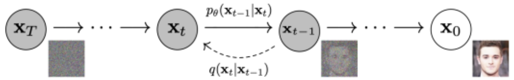
Introduction

Proteins derive their functions from complex three-dimensional structures. These structures, intricately folded based on amino acid sequences, govern proteins' roles in cellular processes.

This poster focuses on a novel method for reconstructing protein structure from contact maps, which are the inter-residue distances of each amino acids

Background

1. Diffusion model


$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Diffusion model is a score-based generative model. It learns a distribution by learning to denoise noisy data of that distribution. The learned distribution is then used for generating a new sample. The components of diffusion model are as follows.

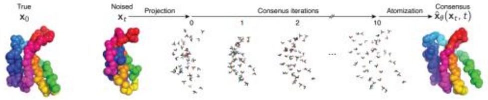
(i). Forward process: this is the noising process that gradually adds noise to data.

$$x_t = \sqrt{\overline{\alpha_t}}\, x_0 + \sqrt{1-\overline{\alpha_t}}\, z$$

(ii). Reverse process: this is the denoising process that gradually restores data from pure noise.

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}|\mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$$

2. Chroma



Chroma is a protein generative model based on diffusion. It uses an SE(3)-equivariant GNN as the noise prediction network (μ_θ). Besides unconditional protein generation, it also implements a variety of conditional generation pipelines. The general method of conditional generation is through projections or by leveraging guidance of external gradients throughout the diffusion process.

Method

We used Chroma, a diffusion model trained on protein structure generation task, as the foundation of our method. Chroma contains knowledge about the inherent properties of protein structures, which constitutes a powerful and accurate prior. This prior information is used to reconstruct a protein satisfying a given contact map, while still obeying biological restrictions such as folding pattern, substructure pattern etc.

Mathematically, our method can be written as:

$$\log p(x|y) = \log p(x) + \log p(y|x) + C,$$

where $x \in \mathbb{R}^{4N\times 3}$ is the backbone atom coordinates, $y \in \mathbb{R}^{N\times N\times 4}$ is the atom-wise contact map, and C is a constant irrelevant to x.

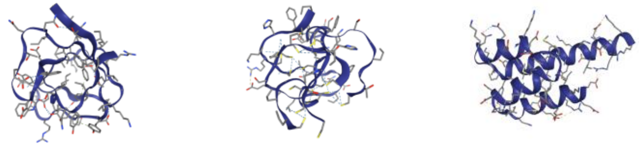
In the recovery process, we do generation using conditional reverse process:

$$\nabla_x \log p(x_t|y) = \nabla_x \log p(x_t) + \nabla_x \log p(y|x_t),$$

where x_t is the noisy coordinates in the reverse diffusion process, and y is the contact map. $\log p(y|x_t)$ is defined as the distance between y and contact map of x_t:

$$\log p(y|x_t) = \|y - y(x_t)\|_1$$

The gradient, $\nabla_x \log p(y|x_t)$, is taken w.r.t. x_t, so that optimizing along this gradient guides x_t closer to y. For noisy contact map $\tilde{y} = y + \epsilon$, we can use diffusion posterior sampling to approximately guide the generation process.



(a) ground truth (b) recovered (c) random

Figure 1. Visualization of recovered structure. (b) is recovered from the contact map of (a). (c) is a randomly generated protein with same number of residues.

Evaluation

We propose several metrics to measure the effectiveness of our model:

1. Contact map MAE: mean absolute difference between contact maps.
2. RMSD: root mean squared distance between atom coordinates of target protein structure and recovered protein.
3. TM-align score: another structure similarity matric which is more robust.
4. Sequence identity: similarity between the amino acid sequences of target protein and recovered protein.

We measure these metrics along different target protein lengths from 10 to 200. For each protein length, we generate 10 samples for contact map and unconditional generation respectively.

Since our method is based on a stochastic generative model, we can measure the best recovery out of several trials. Fig. 3 shows the highest TM-score and smallest reconstruction errors out of 10 trials.

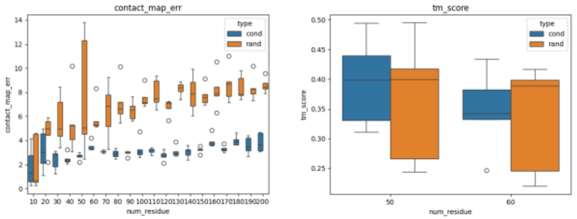


Figure 2. Generated protein with contact map guidance achieves a much lower contact map MAE and a higher TM-score.

Conclusion

Our contact map guided protein structure generation presents a novel way to reconstruct protein structure from contact map, while respecting the inherent constraints of protein structure. The weight of our conditioner is a tunable hyperparameter, and the sampling strategies can also be further tuned for future work.