2019 Spring @ https://qdata.github.io/deep2Read/

Generative Modeling for Protein Structures

Namrata Anand & Possu Huang

Bioengineering Department
Stanford University

ICLR 2018

Introduction

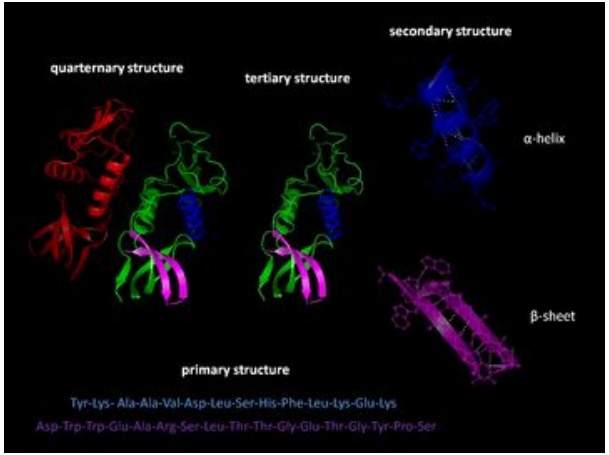
- Task: Computational protein design
- Application: Predicting missing sections of corrupted protein structures
- Key part of understanding biology
- Previous progress in protein design leads to new
  - Therapies
  - Enzymes
  - Small-molecule binders
  - biosensors

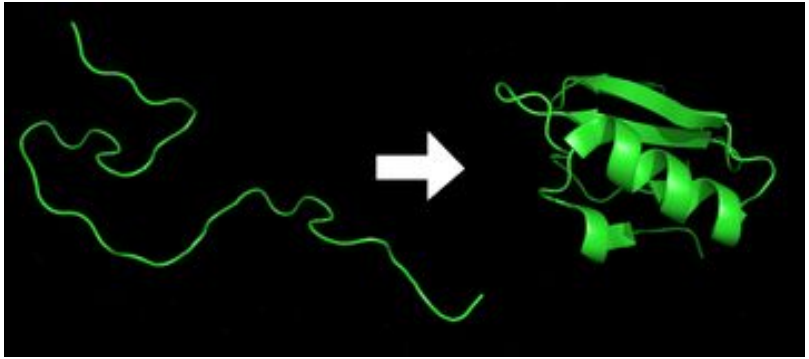
Related Work

- Computational protein/enzyme/small-molecule design
- GAN training stability
  - Martin Arjovsky, Soumith Chintala, and Le´on Bottou. Wasserstein gan. 2017. Link
  - Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. 2016. Link
- Image inpainting
  - Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. 2016. Link

Proteins

- Chains of amino acids
- Chain forms the “protein backbone”
- Structure from backbone folding
- Folding essential for function





Task

- Two parts to designing proteins:
  - Design the scaffold of a structure
  - Find the sequence of amino acids to fold into the structure
- Rosetta technique (past)
  - Samples from native protein fragments for folding backbones
  - Optimizes energy function for most likely orientations and amino acid sequences
- Focus on first part - generate and design new protein structures

Contributions

1. Generative model for proteins that is invariant to rotational and translational symmetry
2. Differentiable, robust complex formulation to solve protein impainting
    a. Can be extended to any problem for structure recovery from pairwise distance measurements

Methods

- Generating Maps
  - Data representation of 3-D Structures
  - 2-D pairwise distances (maps) between alpha carbons on the backbone
  - DCGAN trained on length-matched fragments
  - Generated 16-, 64-, 128-, and 256-residue (monomer/amino acid) maps

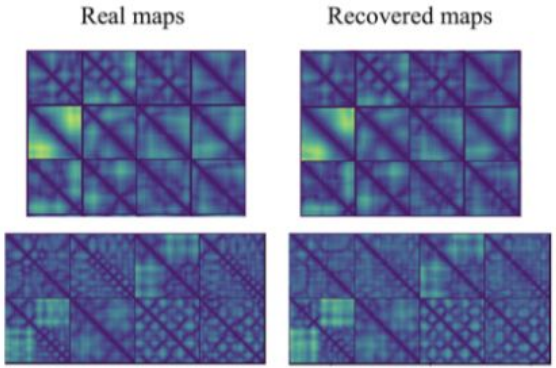


Figure 4: Recovery of maps for 64-residue (top) and 128-residue (bottom) models by optimization of GAN input vector **z**

Methods

- Folding maps via ADMM (alternating direction method of multipliers)
  - Solves convex optimization more efficient than Rosetta
  - Way of **retrieving 3-D cartesian coordinates given pairwise distance measurements**
  - iteratively project updated gram matrix onto the cone of symmetric PSD matrices of rank 3

$$G_{k+1}, \eta_{k+1} = \underset{G,\eta}{\operatorname{argmin}} \left[\lambda \|\eta\|_1 + \frac{1}{2} \left(\sum_{i=1,j=1}^{m} (g_{ii} + g_{jj} - 2g_{ij} + \eta_{ij} - d_{ij}^2)^2 \right) + \frac{\rho}{2} \|G - Z_k + U_k\|_2^2 \right]$$

$$Z_{k+1} = \Pi_{\mathcal{S}_+^m}(G_{k+1} + U_k)$$

$$U_{k+1} = U_k + G_{k+1} - Z_{k+1}$$

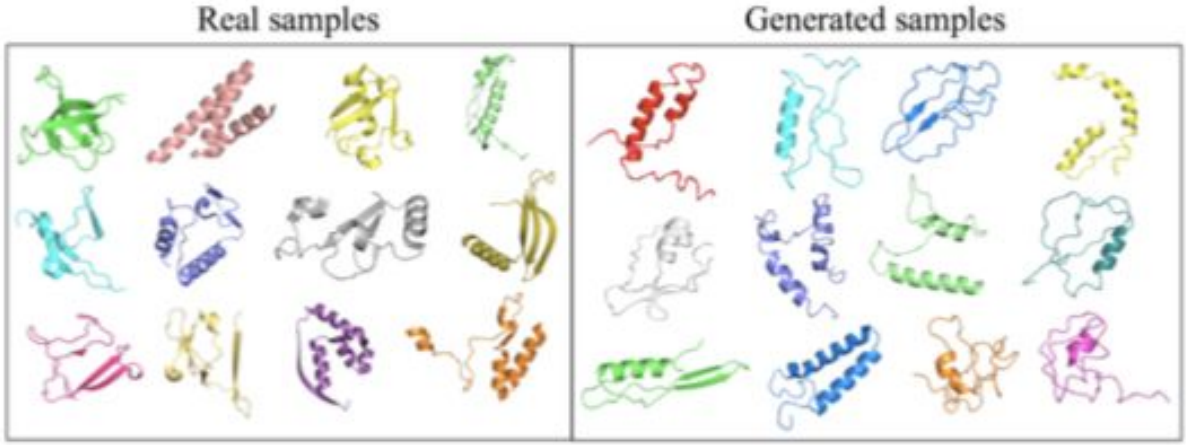


Figure 3: Examples of real (left) 64-residue fragments from the training dataset versus generated (right) 64-residue fragments folded subject to distance constraints using Rosetta.

Inpainting Problem

Notation:
- binary mask M
- mask complement M^C
- input x
- weights W

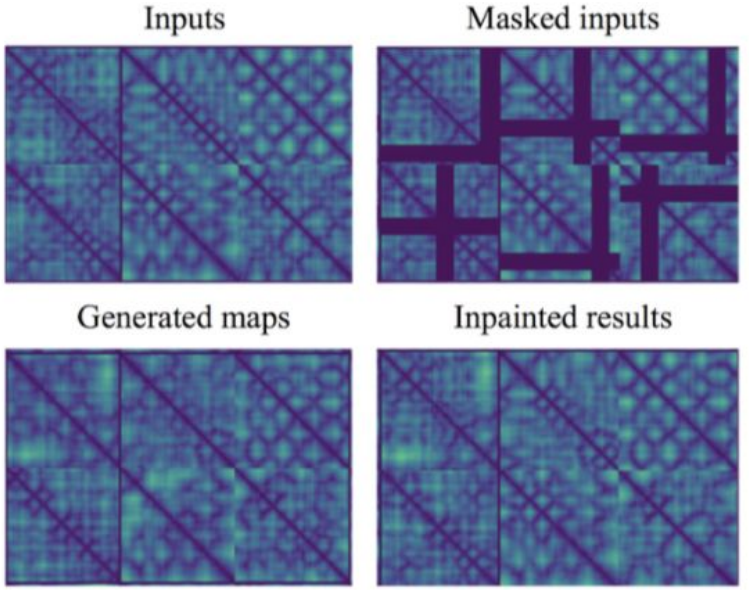


Figure 2: Examples of inpainting for 20 missing residues on 128-residue maps.

Inpainting Problem

- Context Loss $L_{\text{context}}(\mathbf{z}) = \|(W * M^C) * (G(\mathbf{z}) - \mathbf{x})\|_1$
- Prior Discriminator Loss $L_{\text{disc}}(\mathbf{z}) = \log(1 - D(M * G(\mathbf{z}) + M^C * \mathbf{x}))$
- Discriminator loss on final solution $L_{\text{prior}}(\mathbf{z}) = \log(1 - D(G(\mathbf{z})))$
- Full objective:

$$\min_{\mathbf{z}} \; L_{\text{context}}(\mathbf{z}) + \gamma \, L_{\text{prior}}(\mathbf{z}) + L_{\text{disc}}(\mathbf{z})$$

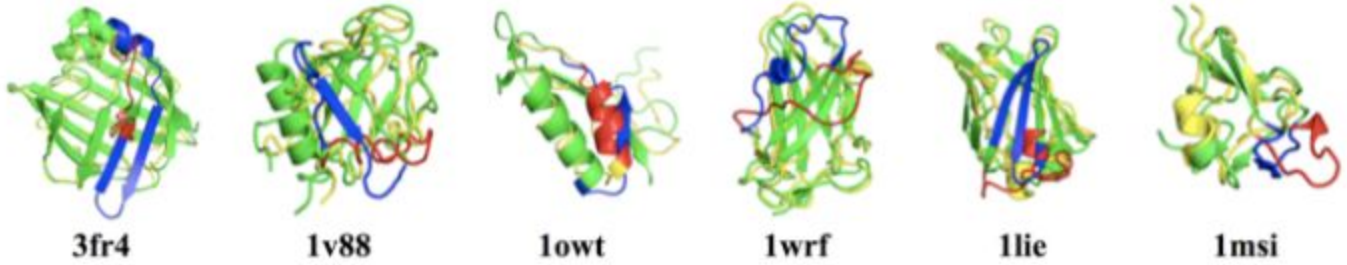


Figure 5: Examples of non-native and incorrect 20-residue (left) and 10-residue (right) inpainting solutions for selected 128-residue and 64-residue structures, respectively, folded using ADMM (PDB ID listed under structure). Native structures are colored green and reconstructed structures are colored yellow. The omitted regions of each native structure are colored blue, and the inpainted solutions are colored red.

Works Cited

Protein Folding: https://en.wikipedia.org/wiki/Protein_folding

OpenReview comments:https://openreview.net/forum?id=HJFXnYJvG

ADMM: http://stanford.edu/~boyd/admm.html

Paper: https://openreview.net/pdf?id=HJFXnYJvG