BetaFold: A Lightweight Protein Prediction Architecture

Sheikh Shams Azam
3rd Year Ph.D. Student
Elmore Family School of ECE
Purdue University

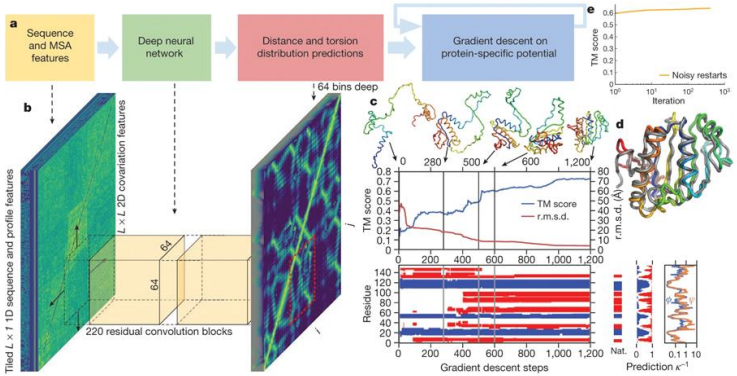
Project Presentation for ECE 695 BH
Apr. 20, 2022

Outline

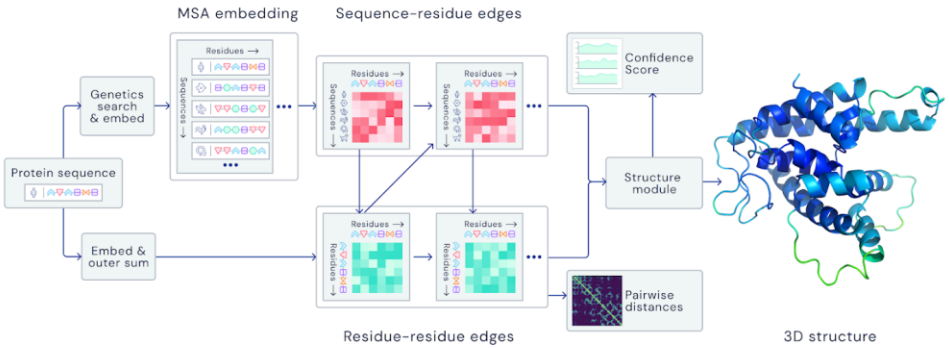
- Problem Definition
- Method
- Implementation
- Evaluation
- Conclusion

Problem Definition

AlphaFold I



AlphaFold II



Bottlenecks:

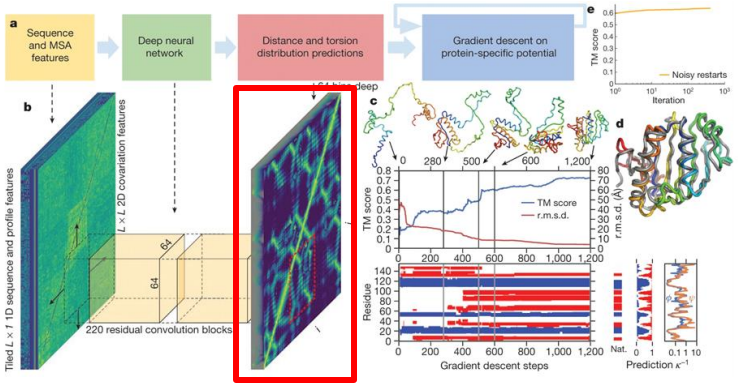
- Open-source implementation for running the inference and not the training
- Released code does not help determine the preprocessing.
- The inference per amino-acid sequence is around 10 mins on a 8 GB Quattro RTX 400, 128 GB machine.
- A major portion of inference is due to Multi-Sequence Alignment (MSA).
- Total size of the original data is around 2TB and ore-training time is over 2 weeks.

Method

BetaFold as a step towards solution:

- Breakdown Alphafold into its constituent and concentrate on Evoformer
- Skip Alphafold preprocessing (not available openly) instead use preprocessing as mentioned in [1].
- Skip MSA prediction: harder to handle 400 channel input feature on a small GPU.
- Focus on distogram prediction as it can be handled as an “image”-to-“image” translation task.

AlphaFold I



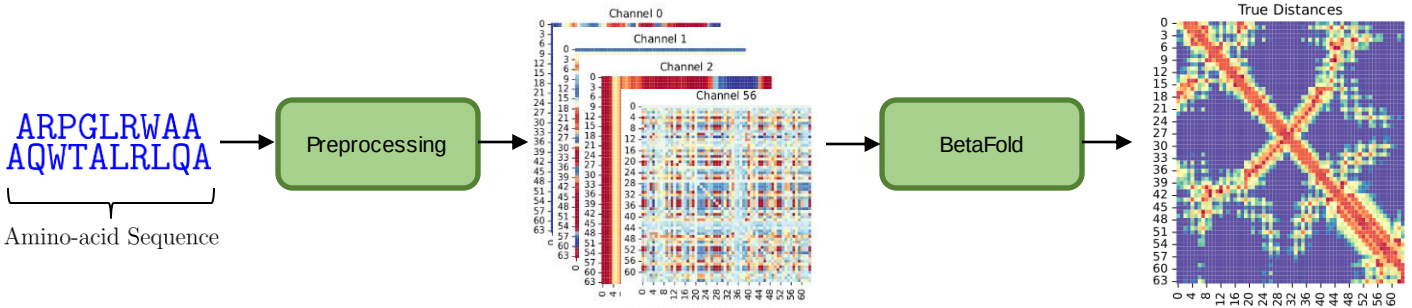
AlphaFold II

1) Jones, D. T. and Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural netw orks and minimal sequence features. Bioinformatics, 34(19):3308–3315, 2018.

Method

BetaFold as a step towards solution:

- Amino-acid sequence preprocessing as proposed in [1].
- Distogram prediction as an “image”-to-“image” translation task.

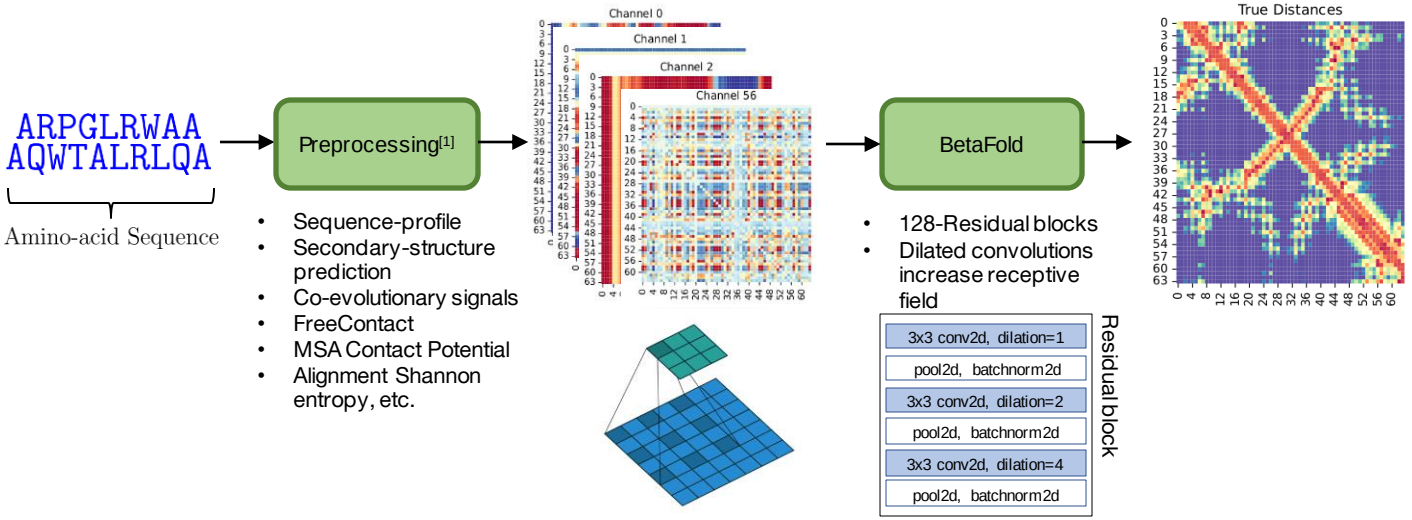


1) Jones, D. T. and Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural netw orks and minimal sequence features. Bioinformatics, 34(19):3308–3315, 2018.

Implementation

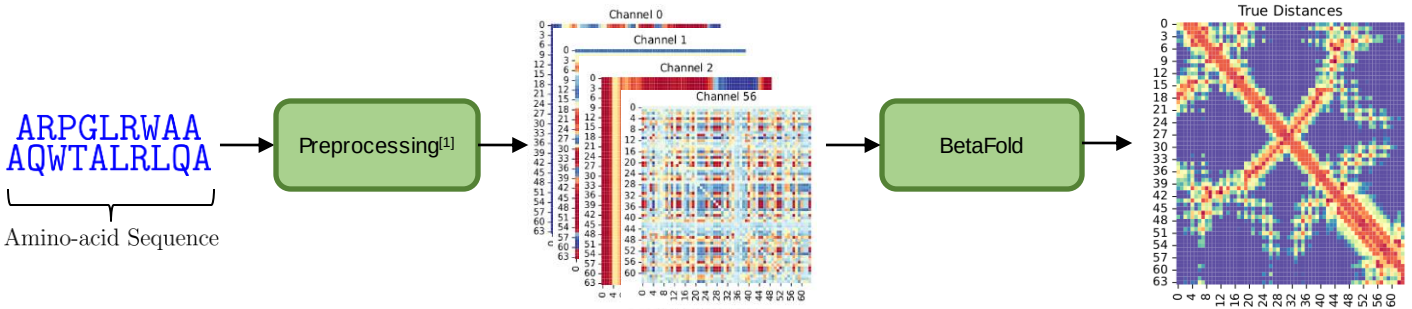
BetaFold as a step towards solution:

- Amino-acid sequence preprocessing as proposed in [1].
- Distogram prediction as an “image”-to-“image” translation task.



1) Jones, D. T. and Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural netw orks and minimal sequence features. Bioinformatics, 34(19):3308–3315, 2018.

Implementation



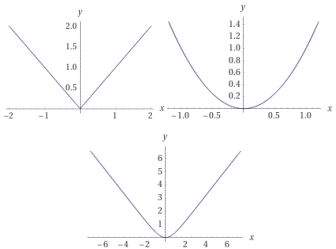
- Two types of task: (i) contact prediction, (ii) true distance prediction
- Contact prediction: binary classification with threshold of 6 angstrom using binary cross-entropy loss
- True distance prediction: using log_cosh and inv_log_cosh, s.t.,

$$\mathtt{log\_cosh}(\mathrm{y}, \hat{\mathrm{y}}) = \log(\cosh(\hat{\mathrm{y}} - \mathrm{y}))$$

$$\mathtt{inv\_log\_cosh}(\mathrm{y}, \hat{\mathrm{y}}) = \log\left[\cosh\left(\frac{\mathrm{K}}{\hat{\mathrm{y}} + \varepsilon} - \frac{\mathrm{K}}{\mathrm{y} + \varepsilon}\right)\right]$$

works like MSE for small values and MAE for large values (i.e., not strongly affected by outliers)

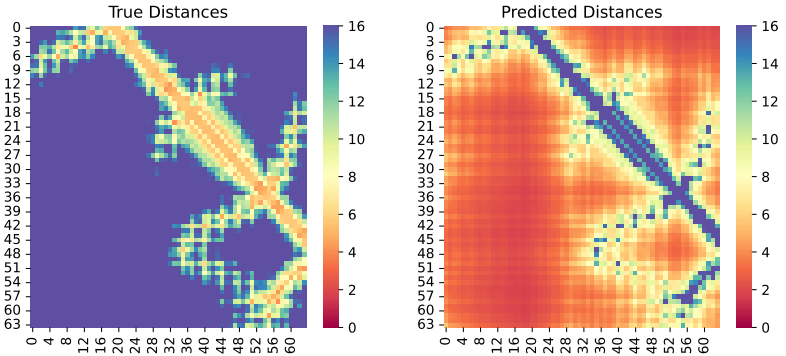
- In practice log-cosh works better but with inverted distances (K/y).
- Implementation at: https://github.com/shams-sam/BetaFold



1) Jones, D. T. and Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural netw orks and minimal sequence features. Bioinformatics, 34(19):3308–3315, 2018.

Evaluation

#	Task	Hyper-parameters	Metric	Performance
1	True Distance Pred.	$\texttt{loss} = \texttt{inv\_log\_cosh}, \texttt{lr} = \texttt{1e} - 2$	MAE	0.27
2	True Distance Pred.	$\texttt{loss} = \texttt{inv\_log\_cosh}, \texttt{lr} = \texttt{1e} - 3$	MAE	0.14
3	True Distance Pred.	$\texttt{log\_cosh}, \texttt{lr} = \texttt{1e} - 2$	MAE	0.30
4	True Distance Pred.	$\texttt{log\_cosh}, \texttt{lr} = \texttt{1e} - 3$	MAE	0.31
5	True Distance Pred.	$\texttt{log\_cosh}, \texttt{lr} = \texttt{1e} - 2$ with inverted distances	MAE	0.08
6	True Distance Pred.	$\texttt{log\_cosh}, \texttt{lr} = \texttt{1e} - 3$ with inverted distances	MAE	0.08
7	Contact Pred.	$\texttt{log\_cosh}, \texttt{lr} = \texttt{1e} - 2$	Accuracy	0.93
8	Contact Pred.	$\texttt{log\_cosh}, \texttt{lr} = \texttt{1e} - 3$	Accuracy	0.92



Conclusion

- AlphaFold is a highly specialized architecture with huge networks trained on really large dataset.
- Lightweight architectures are only possible if we can think of solutions around shrinking the MSA representations.
- Possibly use domain knowledge to bake more inductive priors within the network. Currently attention does the work.

- We try to emulate the task of Evoformer as an “image”-to-“image” translation task.
- Possible application of manifold learning in predicting the 3D structures from 2D distograms.

1) Jones, D. T. and Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural netw orks and minimal sequence features. Bioinformatics, 34(19):3308–3315, 2018.

Thank you!

Any questions/feedback?

- Email: azam1@purdue.edu

References

1. Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
2. Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." *Nature* 577.7792 (2020): 706-710.