In Silico Protein Evolution

Kiana Seraj

July 1, 2024

1 Introduction

Numerous progress has been made in protein structure prediction through the application of deep neural networks. Models like ESM (Evolutionary Scale Modeling) and AlphaFold have demonstrated impressive capabilities in accurately predicting protein structures. ESM, as a language model trained on natural sequences only, exhibits faster feature prediction compared to models that consider both sequences and Multiple Sequence Alignments (MSA), such as AlphaFold [2]. Therefore, in here it has been attempted to examine ESM to see if its captured features can be employed to simulate sequence evolution and generate sequences distinct from natural ones. The method considers both amino acids and backbones in the protein generation process.

The simulation initiates with a population of randomly generated protein sequences, initially resulting in featureless predictions when subjected to the ESM model for structure prediction. Subsequently, a Genetic Algorithm is employed as an evolution algorithm to iteratively optimize the population towards desired structural constraints. This optimization process takes into account protein backbones and evaluates structural scores. Each sequence is assigned a fitness score, and sequences with high fitness scores are selected. The optimization steps involve applying crossover and mutation operations on the sequences.

Optimizing from randomly generated proteins with a 20% dissimilarity in residues yields a population with high mean pLDDT scores. This emphasizes that the ESM model, initially trained on natural proteins, can effectively hallucinate sequences and their backbones with high confidence. This suggests a promising potential for utilizing the model in generating de novo proteins that differ significantly from natural ones [4].

2 Methods

2.1 Hallucination

The Hallucination method has utilized ESMFold to generate backbones prediction, and optimized the intermediate stages iteratively. This leads to generating of protein sequences and their corresponding structures that are not present in the ESM training. This capability is valuable for exploring novel protein structures and sequences that may have unique properties.

2.2 optimization with an Evolutionary Algorithm

One of the most promising application of Genetic Algorithm is data analysis and molecular biology prediction [3]. In here genetic algorithm is used to search in a space of random protein sequences to optimize the sequences toward the favored structural constraint defined in a fitness function. Consequently, the new population suitable for mating will be chosen out of sequences. Each selected protein sequence, having higher fitness score, will be modified with two operators: crossover and mutation.

- Crossover
  Firstly, predicted Aligned Error(PAE), a measure for the confidence in the relative positions and orientations of parts of the predicted structures shown in 2D plot [1], is used to cluster domains of each sequence. Subsequently, the crossover happens by swapping the obtained domains within two sequences.
- Mutation
  Two mutations such as insertion and substitution are done on sequences(insertion is only applied on sequences shorter than a threshold). The substitution is applied with the masked prediction of ESM2, the masked amino acids ,which are 5% of the residues, will be substituted with a selection from 5 most probable amino acids predicted by ESM2.

2.3 Constraint implementation as a Fitness Function

In this section, some structural constraints such as the metrices predicted by esmfold (plddt, ptm, pae) has been used to optimize the sequences toward the sequences having high structural scores. the next fitness scores are defined based on the obtained backbones with the ESMFold. Globularity and translational symmetry for the requiring filaments. Due to the considerable time required for structural prediction of multimers by ESMFold, in here dimers are generated. Then the general fitness score is computed as the linear combination of all constraints each multiplied by a constant as their weight. The more important is the fitness score the bigger is the weight constant. The defined fitness scores are:

- Structure Prediction Confidence
  ESMFold gives a PTM score which is the overall model's confidence in the structure prediction and pLDDT which is per residue model's confidence in atomic coordinate prediction and PAE matrix which is the inter residue relation distances given by the model.

- Globularity
  To have a protein packed into a globular, fitness function which computes the standard deviation of atomic coordinate distances from the protein's centroid has been defined.

- translational symmetry
  In the context of multimeric protein structures, a fitness function has been devised to assess the presence of translational symmetry among consecutive monomers. This symmetry fitness function is characterized by the standard deviation of distances between the centroids of monomers and the corresponding distances along the backbone coordinates. Ideally, this metric goes towards minimal deviation, converging towards a value of zero.

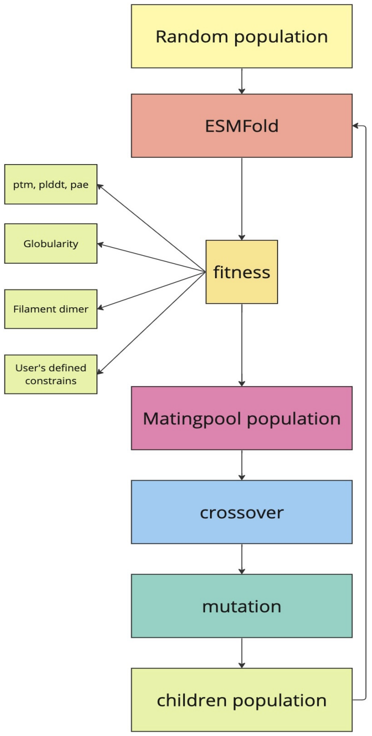

Figure 1: The flowchart of the algorithm

3 Results

After 60 iterations the model reaches the mean plddt of 78 with the variance of less than 0.5 shown in Fig.2a, the generated sequences have 80% of residue similarity with each other. Starting from a featureless random population having 80% similar amino acids , almost linearly, at each step the population's mean plddt increases, the regression plot has been shown in Fig.5

In the generated sequences there has been a more frequent repetition of special amino acids such as K, E, L(Fig.4a,4b) such that the repetition of L demonstrates the formation of hydrophobic core in the protein structure which makes it to be more stable. The repetition of E and K which are charged amino acids, probably forms the surface of the protein.

The attempt to achieve translational symmetry for longer sequences is still in progress. The structural filament dimer depicted in Fig.2b exhibits dimers positioned in front of each other, rather than in a consecutive form. A higher fitness weight for translational symmetry may be more beneficial. Another goal here was also to maintain the length of the sequence population long, which could potentially impact the translational symmetry metric. The translational symmetry for shorter sequences has been achieved showing in Fig.3a, and 3b.

The generated structures so far exclusively exhibit the alpha helix secondary structure. This outcome arises from the local-based formation of the alpha helix, which is more achievable through hallucination. In contrast, the generation of beta sheets involves a different process, relying on the global structure.

The generated proteins were subjected to BLAST analysis to identify similar sequences in natural proteins. However, no matching sequences to these generated proteins were found.

(a) A generated monomer, having the residue length of 195 with the mean plddt of 78.83

(b) The generated dimer of fig 2.a

Figure 2

(a) A generated monomer, having the residue length of 85 with the mean plddt of 74.40

(b) The generated dimer of fig 3.a

Figure 3

(a) A generated domain having respectively the residue length and mean plddt of 34, 84.58

(b) A generated domain having respectively the residue length and mean plddt of 39, 77.34

Figure 4

4 Conclusions

The generated sequences had no similar families on BLAST which demonstrates the capability of language models, trained on natural proteins, to produce novel proteins that diverge from natural ones. The hallucinated population achieved the structural confidence in just 60 steps from a random protein space, highlighting firstly the potential of this method to be so much faster than other optimization algorithms with thousands of iterations, and secondly the potential of hallucination techniques in generating proteins with distinct characteristics.

5 Code availability

The python code used to generate the hallucinated proteins described in the report is available on :
(https://github.com/kianaseraj/ProteinEvolution)

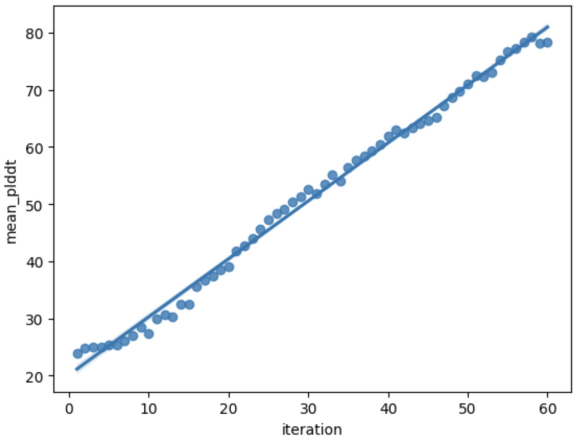

Figure 5: mean plddt of the filament population at each step

References

[1] Christoph Elfmann and Jörg Stülke. Pae viewer: a webserver for the interactive visualization of the predicted aligned error for multimer structure predictions and crosslinks. *bioRxiv*, pages 2023–03, 2023.

[2] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[3] Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[4] Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, pages 2022–12, 2022.