Generation of Protein Sequences with a Given Characteristic

Author: Hugo Hrbáň
Supervisor: doc. RNDr. David Hoksza, Ph.D.

Introduction

Prompt engineering meets protein engineering!

Proteins are biological macromolecules that are crucial for all life on Earth. A protein consists of amino acids, which chemically bond to form long chains that fold into a 3D structure. These protein sequences can be considered sentences in the protein language, which consists of 20 unique amino acids. Thanks to this, we can train **language models** on large databases of protein sequences to learn the properties of the protein language and use these models to generate novel proteins. The task of *de novo* protein design has numerous applications for biological research and development of new drugs to prevent and treat diseases.

Goals

The main goals of this work were to:

- Finetune a generative protein language model based on the Transformer decoder architecture to allow for more controllable generation of protein sequences from selected protein families.
- Evaluate the quality of generated sequences using hidden Markov models (HMM), sequence alignment, and protein structure prediction models.
- Explore the learned representations of the model (attention mechanism, token embeddings).
- Create a pipeline for users to replicate this process on their custom set of protein sequences.

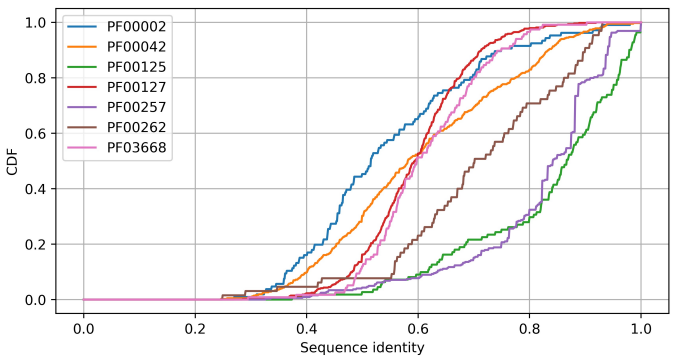
Results

Percentage of generated proteins belonging to each family, detected by profile HMMs.

Profile HMM for family

Generated for family	PF00002	PF00042	PF00125	PF00127	PF00257	PF00262	PF03668
PF00002	71.88	0.0	1.56	0.78	0.78	0.0	0.0
PF00042	0.0	79.88	0.0	0.2	0.0	0.0	1.17
PF00125	0.0	0.0	83.59	0.0	0.0	1.56	0.0
PF00127	0.39	0.2	0.0	92.58	0.0	0.0	0.0
PF00257	0.59	0.0	9.96	0.78	60.94	0.39	0.39
PF00262	1.56	0.78	14.84	4.69	0.78	40.62	0.0
PF03668	0.0	0.78	0.0	0.0	0.0	0.0	89.06

Cumulative distribution of pairwise sequence identity to the finetuning dataset.



Conclusion

We show that pre-trained protein language models can be finetuned to provide a way for users to control the properties of generated sequences, and that they exhibit good *in silico* qualities.

Future work:

- Scaling up model size, datasets and prompt options
- Diffusion models for sequence infilling

Implementation

Download *7* protein families from the Pfam database. (*89k* sequences = *27M* tokens)

Preprocess data: Cluster at *90%* identity, prepend sequences with special family tokens and initialize new rows in the model's embedding matrix.

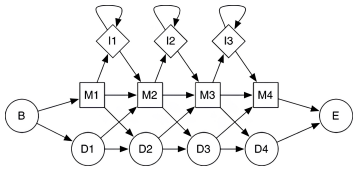
```
<|pf00125|>2PGFDEREGEADADGEPDHPRYAQNPDGAGTPDPGRQ...
<|pf00071|>1MGIANYNKVNGLKAKASKTSLKDKSFESCLDFKSPQ...
<|pf00042|>1RAALGMKAALEVLNMDIQKEFGIQIEVGIGVHFGPV...
```

Finetune the *ProGen2* model on new data for 5 epochs on A100 80GB GPU for *~8 hrs*. We train a one-directional and bidirectional model.

Inference. Generate protein sequences using the model, specify only the family token as the **prompt**. We use top-*k* sampling with temperature.

Evaluation

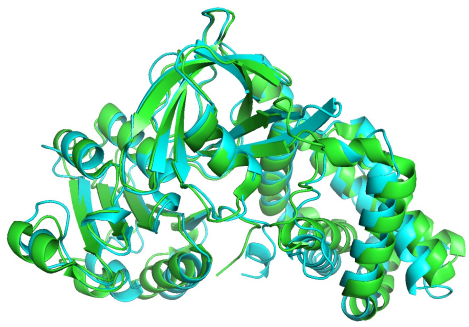
- Use profile HMM to detect if generated sequences belong to correct family
- Align generated sequences with training data and compute percentage of identical residues to see how the model generalizes
- Explore how values of *k* and *temperature* influence properties of generated proteins
- Compare performance of one-directional and bidirectional models
- Detect sequence motifs using regular expressions




```
I..LFLEA.SDSTIIRRYK.ETRRRHP....
V..LYLDA.DEETLLKRFS.ETRRRHP....
S..VFLDA.NTDTLVRRFS.ETRRRHP....
V..LFLDA.SDEALVRRYS.ETRRRHP....
I..VFLDC.SDDTLVARYK.ETRRLPP....
M..LFLEA.SPETVIKRYK.ETRRKHP....
```

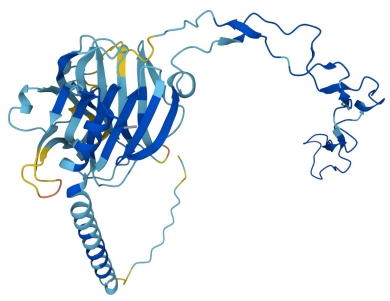
Structure Prediction

We use **ESMfold** to predict structures of generated sequences.

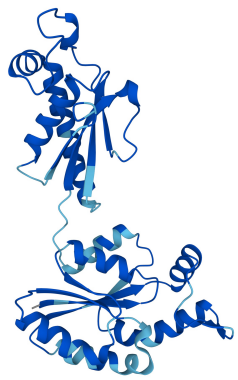


Predicted structure from a sequence generated with prompt `<|PF00042|>1` (blue) aligned with experimentally determined structure 1GVH (green)

Sequence identity: *41.7%*
pLDDT: *90.4*



Generated for family *PF00262*
Sequence id.: *62.9%*
pLDDT: *81.9*



Generated for family *PF03668*
Sequence id.: *58.6%*
pLDDT: *90.9*

Contact: hugohrban2@gmail.com

https://github.com/hugohrban/ProGen2-finetuning

https://huggingface.co/hugohrban