FBGAN を用いた合成タンパク質生成の品質向上

中野 花恋 (指導教員：オベル加藤 ナタナエル)

1　はじめに

合成生物学は製造業，医療，農業，環境問題など様々な分野に飛躍的な進化をもたらしうる学問領域である．現在までの合成生物学は手作業での研究が主であったが，研究対象としている遺伝子という大規模なデータは機械学習に適していると言える．本研究では，畳み込みニューラルネットワーク (CNN) を用いた生成モデルである，敵対的生成ネットワーク (GAN) を合成タンパク質生成に適用させた，Feedback GAN (FBGAN) に着目し，構造を多段階化することで合成タンパク質の品質向上を目指す．

1.1　GAN の概要

GAN[1] とは，Goodfellow らによって提唱された機械学習モデルである．一般的な GAN モデルは，生成器 (Generator) と識別器 (Discriminator) という 2 つの独立したネットワークで構成され，両者が敵対しながら精度を高め合うことでノイズからより本物に近い目標を生成する．

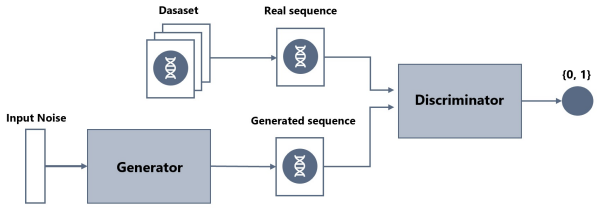

図 1: GAN の構造

1.2　FBGAN の概要

Feedback GAN (FBGAN)[2] とは，GAN に転移学習を組み込むことで，Discriminator の微分可否に関わらず臨んだ適性を持つ遺伝子配列を生成する学習モデルである．FBGAN の最大の特徴は，図 2 に示すように，GAN 構造と Analyzer がフィードバック機構で接続されている点である．Generator が生成した生成データはまず Analyzer に入力され，Analyzer は入力に対してスコアを返す．そのスコアに基づいて Discriminator の本物データが繰り返し置き換わるため，目的の機能を有する配列の生成が可能になっている．

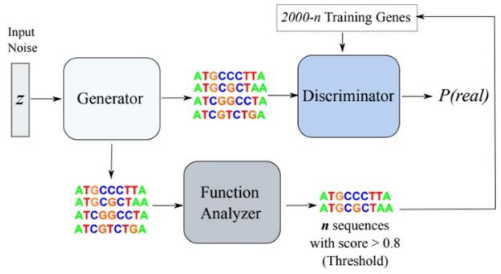

図 2: FBGAN の構造，[2]，4 ページ，Figure 1 より引用

2　関連研究

GAN の仕組みを応用させた研究例として Zhang らの StackGAN[3] が挙げられる．これは，文章を入力として受け取り，それに適した画像を生成する Text-to-Image と呼ばれるモデルの実装である．

StackGAN の特徴の一つとして，まずテキストから低解像度画像を生成し，生成された低解像度画像を基に高解像度画像を生成する手法を採用しており，この部分に着目して GAN による高解像度画像生成の研究例としても注目されている．本研究ではこの多段階構造をノイズからの合成タンパク質生成に応用した．

3　提案手法

従来の FBGAN の場合，Analyzer と接続されている GAN 構造は図 1 に示したようにノイズから直接配列を生成している．そのため，より残基数の多い配列を生成しようとすると多くの時間がかかる．そこで，本研究では図 3 に示すような構造を用いることで同一 epoch 数でより長く複雑な配列の生成を目指す．

3.1　概要と目標

生成器と識別器のセットを 2 つ使用する．両者は直列に接続され，1 つ目のステージでノイズから生成した配列を，2 つ目のステージでは入力として受け取る．2 つ目のステージでは受け取った短い配列を基に，より長い配列を生成することで実行時間を抑制したまま最終的に FBGAN によって生成される配列の質を向上させる．(図 3)

具体的な目標としては，1 段階目では，先行研究と同様に約 50 残基のタンパク質を生成するような合成配列の生成を行う．そして 2 段階目では，先行研究の 2 倍の配列長である，約 100 残基のタンパク質を生成するような合成配列の生成を目指す．

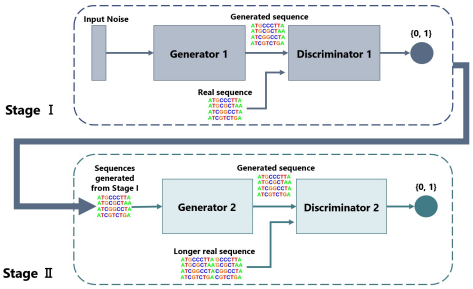

図 3: 本研究の GAN の構造

3.2　データセット

約 100 残基のタンパク質を生成するためのデータセットとして，配列長が 156 から 306 の DNA 配列を用意した．まず，Uniprot データベースが提供するレビュー済みタンパク質のデータセットから，長さが 50-100 残基のタンパク質を収集した．[4] 次に，これらのタンパク質を配列類似度 0.6 以上でクラスタリングし，各

クラスタから代表的な配列を 1 つずつ選択した．最後に，これらの配列を DNA に翻訳することで学習に必要なデータセットを得た．

3.3 実験設定

Stage I：従来の GAN と同様の構造．Generator1 はノイズを入力として受け取り，先行研究と同様に約 50 残基のタンパク質を生成するような合成配列を生成する．ここで Discriminator1 が参照する本物データは，先行研究に用いられていたデータセットである．
Stage II：Stage I で生成した配列を Generator2 の入力とし，約 100 残基のタンパク質を生成するような合成配列を生成する．ここで Discriminator2 が参照する本物データは，本研究で作成された長い配列のデータセットである．

4 実験結果

4.1 配列長

提案した構造を用いて生成した合成配列の長さを先行研究と比較すると以下のようになる．図 4 はすべて同一 epoch 数で学習を進めた結果を [平均値 ± 標準偏差] という形で表している．左列と右 2 列を比較すると，GAN に与えるデータセットを変えることで，生成される合成配列の長さが大きく伸びていることがわかる．また，中央と右列を比較すると，使用しているデータセットは同じでも本研究の提案手法による結果は従来の FBGAN による結果と比較して，より長い配列を生成できていることがわかる．さらに，誤差に着目すると，提案手法は従来の FBGAN よりも極めて安定的に大規模な合成配列を生成できていると言える．

FBGAN	FBGAN + 本研究のデータセット	**本研究**
90.61 ± 14.74	229.16 ± 32.01	**296.53 ± 0.88**

図 4: 実験により生成された合成配列の配列長比較
左: 先行研究の FBGAN による生成結果，
中央: 先行研究の FBGAN に今回作成したデータセットを適用させた生成結果，
右: 本研究の提案手法による生成結果．

4.2 生成精度

提案手法により生成した合成配列が天然配列とどの程度近い特性を持つのかを t-SNE を用いて比較した結果を図 5 に示す．青い点は本研究の提案手法による合成配列，赤い点は本研究で使用した学習データセットから取り出した，Uniprot データベースに存在している天然配列を表している．グラフより，提案手法を用いた構造で生成された合成配列は天然配列と同じ空間に存在していることがわかる．つまり，天然配列の持つ特性に近しい合成配列の生成に成功していると言える．

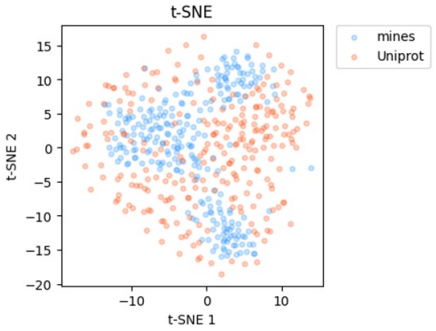

図 5: t-SNE による比較結果

5 まとめと今後の課題

本研究では GAN の構造を多段階化することで，従来の FBGAN よりも長く複雑な合成配列を精度や実行時間を損なうことなく，安定的に生成することに成功した．今後は更に段階を重ねた GAN 構造を用いた場合との結果の比較や，図 6 のように，Analyzer として任意の配列データベースと比較して配列相同性を算出することのできる，HMMER [5] を用いた構造を実装することで，更に天然配列に近しい合成配列を生成する試みを行いたい．

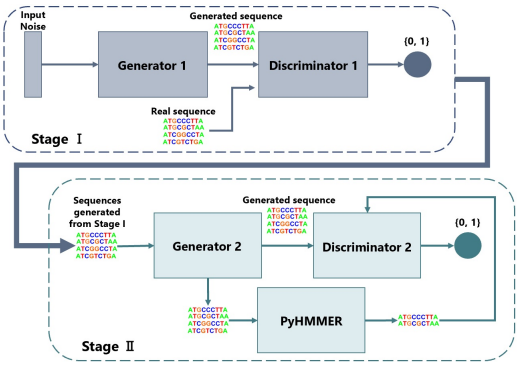

図 6: HMMER の Python モジュールである，PyHMMER[1]を Analyzer として適用させた構造の提案

参考文献

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, Vol. 63, No. 11, pp. 139–144, 2020.

[2] Anvita Gupta and James Zou. Feedback gan (fbgan) for dna: A novel feedback-loop architecture for optimizing protein functions. *arXiv preprint arXiv:1804.01694*, 2018.

[3] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

[4] Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, Vol. 51, No. D1, pp. D523–D531, 2023.

[5] Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, and Robert D Finn. Hmmer web server: 2018 update. *Nucleic acids research*, Vol. 46, No. W1, pp. W200–W204, 2018.

[1] https://github.com/althonos/pyhmmer.git