GAN を用いた情報保存用 DNA 塩基配列の識別及び合成

奥村 真由 (指導教員：オベル加藤ナタナエル)

1　はじめに

現在, 情報技術の急速な発展に伴う情報量の急増により, 膨大な情報やデータの存在が問題となっている. そこで, 新たにデータ記憶媒体として注目されているのが DNA である.
従来のストレージメディアは, 物理的なスペースの浪費, 高い保守コスト, および腐敗しやすい材料という欠点を持つのに対し, DNA はたった 4 つの塩基から生成されているにもかかわらず, 大容量, 高密度, および長期的な安定性という利点を持ち, 新しいタイプのストレージメディアとして広く認識されている.
しかし， DNA に情報を保存することにはデメリットもある. DNA 配列にコード化されたファイルの読み書きには, 大きな時間コストを要する. また, DNA 塩基配列には様々な条件があり, 適切な大規模配列を生成することが難しいとされている.
本研究では, GAN を用いて, 遺伝情報ではなく情報をエンコードするための DNA 塩基配列の識別及び合成を試みる. GAN を用いることで, 直接的な配列の最適化よりも時間コストを抑え, 適切な大規模 DNA 塩基配列の生成に繋がると考える.

1.1　GAN の概要

GAN [1] とは， Goodfellow らによって提唱された,Generative Adversarial Network（敵対的生成ネットワーク）と呼ばれる, 機械学習モデルである. GAN の基本的な構成は, Generator(生成ネットワーク）と Discriminator(識別ネットワーク）の 2 つのコンポーネントネットワークから成り立つ. ここで, Generator はノイズ z から新しいデータを生成し, Discriminator はそのデータが本物か偽物かを判別する.

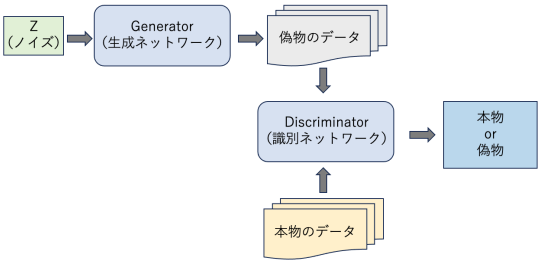

図 1: GAN の構造

Generator と Discriminator は, 式 (1) の損失関数で最小最大ゲームを行う.

$$minmaxV(D,G) = E_{x\in P_{data(x)}}[log(D(x)] + E_{z\in P_{(z)}}[log(1-D(G(z))] \quad (1)$$

本研究では, Wasserstein GAN（WGAN）[2] と呼ばれる GAN の変種を使用した. WGAN は, 一般的な GAN よりもトレーニング中に安定していることが示されている.

2　提案方法

DNA に保存された情報をエンコードするにあたり, エラーを減少させるためには, 適切な制約付きのエンコーディングセットを構築することが必要である [3]. 今回学習データとして, 以下の制約を持ち, 情報保存に適切な DNA 塩基配列（配列長：46〜56、データ数：50）を直接的に合成し、GAN で学習を試みた.

2.1　GC コンテンツ制約

GC コンテンツ制約は, DNA 塩基配列の熱的安定性に関連する. GC と AT のコンテンツの割合がバランスを欠いたオリゴヌクレオチドは, 高い脱離率を持ち, ポリメラーゼ連鎖反応（PCR）のエラーや, 配列解析プロセスへの影響が生じやすくなる. 本研究では, GC コンテンツの割合が 45〜55 %となるようデータセットを生成した.
式 (2) は配列 r に対する GC コンテンツを計算する式である.

$$GC(r) = \frac{|G| + |C|}{|r|} \quad (2)$$

2.2　ラン長制約

ホモポリマーを含む DNA 塩基配列は, シーケンシングエラーの発生確率を大幅に増加させる可能性がある. この問題を防ぐため, ラン長制約を使用し, 隣接する塩基の繰り返しの発生確率を 0 %にすることで, 挿入および削除エラーの防止を図る.

$$l_i \neq l_{i-1}, i \in [1,n] \quad (3)$$

2.3　自己補完制約

各設計された配列において連続した補完ペアとなる塩基が存在する場合, その配列は非特異的なハイブリダイゼーションを経て, 二次的なヘアピン構造（ステムループ）を形成する. プライマーライブラリとして使用される際に, この構造を持つ配列が発生すると, シーケンシングの失敗やデータの損失などの重大な結果を引き起こす可能性がある.
この問題を防ぐため, 設計された配列内でワトソン・クリックの塩基対が 10 個以上連続して存在する場合, 及び同じ塩基が 3 回以上連続して存在する場合を禁止する, 自己補完制約を提案する.

3　実験結果

3.1　実験結果１

本研究で生成した学習データを用いて, GAN で学習を進めた. 図 2 における, d_fake, d_real は, それぞ

れ, Generator が生成した偽物データ, 本研究で生成した本物データを, Discriminator が”本物”として識別した値である. また, 図 3 における, Loss_discriminator, Loss_generator は, 学習結果の平均値である.

d_real の値が d_fake を上回っていることから, 本物データを ” 本物 ” として, 偽物データを ” 偽物 ” として識別できていると言える. 即ち, 学習に成功していると考えられる.

また d_real 値の増加よりも, d_fake 値の減少が速いことから, 現段階では, Generator が生成した偽物データを ” 偽物 ” として識別する学習が進んでいる.

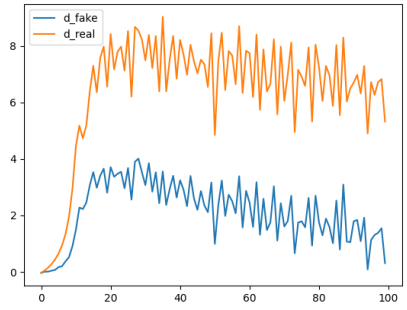

図 2: d-fake, d-real の値の変化

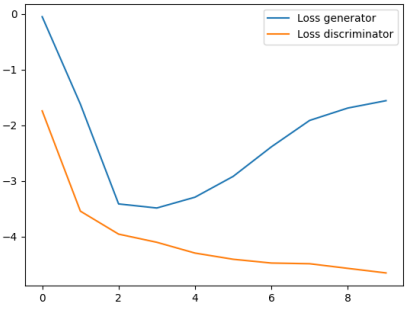

図 3: Loss_discriminator, Loss_generator の値の変化

3.2 実験結果 2

学習を進めた GAN を用いて生成された DNA 塩基配列の, 配列長に関する評価を実施した. 図 4 は, 結果を [平均値 ± 標準偏差] という形で表している. 誤差に着目すると、提案手法によって学習を進めた GAN で生成された配列は, ランダムに生成した DNA 配列よりも, 安定的であると言える. このことから, GAN は, 制約を満たした”本物”に近いデータ, 即ち情報保存における DNA 配列に適切な配列を安定的に生成することが可能であると考えられる.

ランダムデータ	GANから生成された配列
51.04 ± 3.45	54.80 ± 1.02

図 4: 実験により生成された合成配列の配列長比較

'AAGACACAGCACAGACGTTAGATTACGCTACGCTCGTGGTAACATACTACATCCT'
'AACGTAAGTAATAGATGGCGCTCTACAAGGTCGCTGTTGAGTAACGCGGAACG'

図 5: GAN から生成された適切な DNA 配列の例

4 まとめと課題

本研究では, GAN を用いて情報保存のための DNA 塩基配列の識別及び生成に成功した.
今後は, GAN での学習をより進め, 識別及び生成精度を上げる. また, 生成した DNA 塩基配列が適切かどうかを評価する評価関数 (Analyzer) を GAN に組み込んだ Feedback GAN (FBGAN) [4] を用いて, 情報保存により適切な配列を生成できるよう取り組む予定である. 加えて, 保存できる情報量を増やす評価モジュールの生成, 実行を試みる.

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 adversar- ial networks. Communications of the ACM, Vol. 63, No. 11, pp. 139–144, 2020.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[3] Qiang Yin, Yanfen Zheng, Bin Wang , and Qiang Zhang：Design of Constraint Coding Sets for Archive DNA Storage.IEEE/ACM Transactions on Computational Biology and Bioinformatics ,Volume: 19, Issue: 6, 01 Nov.-Dec. 2022. pp.3384-3394.

[4] 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.