QD アルゴリズムを用いた情報保存用 DNA 塩基配列合成の最適化

平田　陽香　(指導教員：Natanael Aubert-Kato)

1　はじめに

情報技術の急速な発展に伴った情報量の急速な増加に伴って情報記録媒体の需要が高まっている現代において、情報密度や保存性に特に長けている新たなデータ保存媒体として注目されているのが DNA である。本研究では QD アルゴリズムの一種である”MAP-Elites”を用い、情報保存用 DNA 塩基配列合成の最適化を図る。

1.1　DNA ストレージの概要

[1] DNA ストレージとは、CD や USB メモリなどにデータを書き込んだり読み込んだりするように、DNA を情報媒体として利用するアイデアのことである。1988 年に Joe Davis によって初めて試みられ、これまで様々な手法が研究されてきた。情報記憶媒体として、DNA は保存性と情報密度の点で極めて優れている。DNA は適切な低温環境下では理想的には約 30 億年間保存できる。理論的には、DNA1 グラムに最大 680PB（6 億 8000 万 GB）、1 立方ミリメートルあたり数 EB（数十億 GB）のデータを保存でき、これは HDD の約 100 万倍の情報密度に相当する。このことから、DNA はほぼ永久にデータを保存できると言われている。

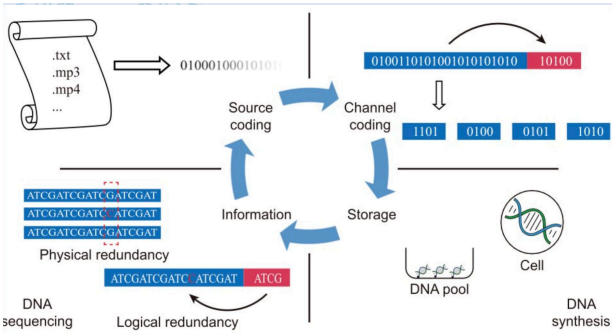

図 1: DNA エンコーディングの流れを表す（左上）ソースコーディング,（右上）チャンネルコーディング,（右下）DNA 塩基の物理的保存,（左下）DNA 塩基配列のデコード [1]

1.2　QD アルゴリズム”MAP-ELITES”の概要

本研究で利用する MAP-Elites とは, Mouret らによって提唱された QD アルゴリズムの一種である。[3]
MAP-Elites は以下のステップで動作する。

1. 初期値としてランダムな解（遺伝子型）を N 個生成する。
2. 生成した解集団を特徴空間のセルにマッピングする。
3. 解集団から無作為に選択した解に突然変異を適用して新たな解を生成する。
4. 新たな解の特徴量と適応度を計算する。
5. その解が属するセルに残してある解と比較して適応度が高ければ入れ替える。
6. ステップ 3〜5 を繰り返す。

従来のアルゴリズムが単一の最適解を見つけることに集中するのに対し、MAP-Elites は探索空間全体における高性能な解を網羅的に見つけ出し、異なる属性を持つ多様な解を生成することを目的とする。また、MAP-Elites は解の「品質」を示す性能だけでなく、その解の特徴や属性の異なる組み合わせが探索空間全体でどのように分布しているかをプロットすることが出来る。探索空間を広く探索するするため計算量が多いという欠点はあるものの、従来の進化的アルゴリズム（EA）や多様性を持つ EA（EA+D）よりも、特徴空間全体を通して高性能な解を見つける能力に長けている。

2　提案手法

本研究では、以下のように評価関数と特徴空間を設定することにより、ランダムに生成した DNA 塩基配列の最適化を試みた。先行研究において確認された以下の知見 [2] を今回の最適化に活用した。

- GC と AT の割合がバランスを欠いた DNA 塩基配列は高い脱離率を持ち, 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。
- ホモポリマー（同じ塩基が連続する配列）の長さに対してインデル (挿入・欠失) 率が上がっていく。
- Watson-Crick 塩基対の繰り返しの長さに対してインデル (挿入・欠失) 率が上がっていく。

2.1　評価関数

2.1.1　50%から離れた GC コンテンツ率に対するペナルティ

塩基列全体を占める GC の割合が 50%から遠いほどペナルティを与えることとした。

2.1.2　同じ塩基の繰り返しに対してのペナルティ

同じ塩基が 4 つ以上連続した場合、その長さに応じてペナルティを与えることとした。

2.1.3　Watson-Crick 塩基対繰り返しに対するペナルティ

DNA 塩基列内で Watson-Crick 塩基対（AT,TA,GC,CG）が 3 個以上連続して存在する場合、その長さに応じてペナルティを与えることとした。

2.2　特徴空間

2.2.1　エントロピー（平均情報量）

配列のエントロピーをを特徴空間の変数の一つとして取り入れた。

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

2.2.2 配列の長さ

配列の長さを特徴空間のもうひとつの変数とした。QD アルゴリズムの中の DNA 塩基配列生成の段階において、配列の長さは 5-20 の間に制限した。

3 実験結果

前述した手法を用いて、各変数に適切な重みづけを行い MAP-Elites アルゴリズムによる学習を進めた。一回あたり 10000 の塩基列を生成し全部で 10 回実施した。計 100000 塩基列分の学習を行うこととなった。

3.1 MAP-Elites によるプロット

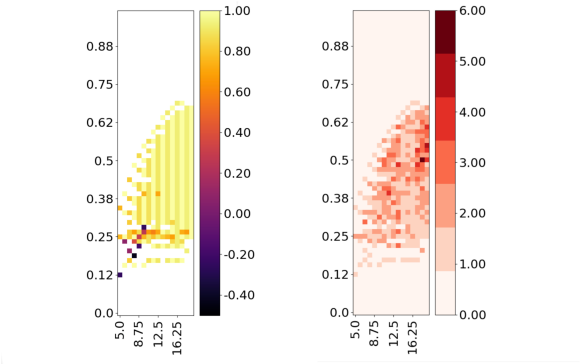

図 2: MAP-Elites の実行結果の例

図 2 の 2 つのマップは、横軸が塩基列の長さ、縦軸がエントロピーを表している。そして、左側のマップでは、各セルに残された個体の適応度の数値に対応した色が表示されている。右側のマップでは、各セルが何回更新されたか対応した色が表示されている。

各実行時にプロットされた個体の適応度の多くは 1.00(最大値) にかなり近いものであった。特徴空間においてある程度満遍なく良い評価が見られるが、配列が短くエントロピーが少ないエリア（プロットの左下）では低い評価となっていた。

3.2 最適化された塩基配列の特徴

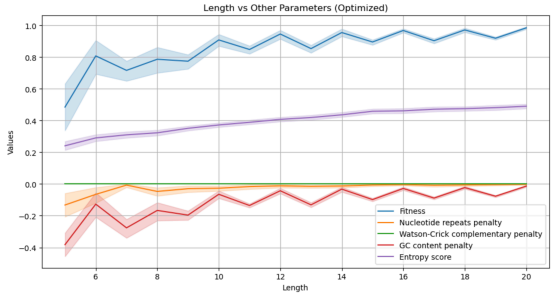

図 3: 本研究における最適化後の塩基列の長さとその他の値の関係 n=100000

10 回の学習を通して最後まで残った計 100000 塩基列に関して、横軸を配列の長さ、縦軸を適応度、同じ塩基の繰り返しに対するペナルティ、Watson-Crick 塩基列の繰り返しに対するペナルティ、GC 率に対するペナルティ、エントロピー、としてプロットを行った。

比較のため、5-20 それぞれの長さにあたり 10000 塩基列ずつ、計 160000 塩基列をランダムに生成した場合にも同様のプロットを行った。

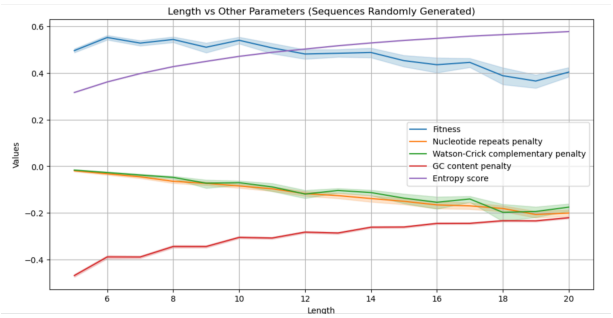

図 4: ランダムに生成した塩基列の長さとその他の値の関係 n=160000

ランダムに生成した塩基列では長さが長いほど同じ塩基の繰り返しや Watson-Crick 塩基列の繰り返しに対するペナルティが大きくなりやすいが、最適化によってこれらのペナルティを減らし、適応度を高めることができている。特に、Watson-Crick 塩基対や同じ塩基の繰り返しに対するペナルティは最適化を経てほとんどすべて 0 になっている。対して、GC の割合に関しては 50%の個体に限りなく近い個体だけでなく、GC の割合に対する評価が-0.4(50%から ± 16%にあたる) である個体も残っていることがわかる。適応度が低くなるような特定の塩基列が最適化を経て捨てられているため配列のエントロピーが少し下がっているが、評価関数に関係する値が飛躍的に向上している。エントロピーが少し下がっても問題はないことがわかる。

4 まとめと今後の課題

まとめと今後の課題本研究の手法で、ランダムに生成した場合と比べてより情報保存に適した DNA 塩基配列へと最適化することに成功した。

本研究においては個々の塩基列に対して最適化を行ったが今後は多くの塩基列の含まれるデータセットを扱う予定である。

参考文献

[1] Dong, Y., Sun, F., Ping, Z., Ouyang, Q. and Qian, L.: DNA storage: research landscape and future prospects, *National Science Review*, Vol. 7, No. 6, pp. 1092–1107 (2020).

[2] Johnson, M. S., Venkataram, S. and Kryazhimskiy, S.: Best practices in designing, sequencing, and identifying random DNA barcodes, *Journal of molecular evolution*, Vol. 91, No. 3, pp. 263–280 (2023).

[3] Mouret, J.-B. and Clune, J.: Illuminating search spaces by mapping elites, *arXiv preprint arXiv:1504.04909* (2015).