機械学習モデルを導入した品質多様性アルゴリズムによる効率的な DNA 構造体探索の実装

理学専攻・情報科学コース g2240655 竹口 葵衣 (指導教員：オベル加藤ナタナエル)

February 3, 2024

1 イントロダクション

分子やゲルを材料とするナノデバイス「分子ロボット」は幅広い分野で応用可能なことで注目されており，特に DNA 分子がなす構造体（以降，DNA 構造体）を部品として用いることが一般的であるが，有用な DNA 構造体や，その合成に必要な塩基配列を特定することは困難である．そのため最適な候補を多様に得る「品質多様性アルゴリズム（QD）」による探索が考えられるが，DNA やその構造体の探索に適用すると膨大な計算負荷が生じる．そこで本研究では，Cazenille ら [1] が開発した QD ベースの構造探索ツールと，ダイナミクスモデルの導入により効率的な解の探索を実現する「DA-QD」をベースとし，求める DNA 分子の組み合わせを効率的に探索することを試みる．

2 関連研究

2.1 MAP-Elites

「QD」とは進化的アルゴリズムの一種で，多様で高性能な解の大規模なコレクションを得ることを目的とする．中でも MAP-Elites は，解を特徴に対応するマップ上の位置に配置していくことでコレクションを得る．ランダムな解が配置されたマップで開始し，そこから無作為に解を選んで変異させ，特徴に応じた位置について，より優れた性能であればマップに配置・上書きすることを繰り返す．その結果，解は特徴記述子ができるだけ多様に，かつ探索基準とした値の性能がより優れるように発見されていく．本研究では，マップはグリッド（座標）として取得できる．後述の構造探索ツールも MAP-Elites をベースとしている．

2.2 Cazenille らの構造探索ツール

本研究では Cazenille ら [1] の構造探索ツールを用いてストランドセットの情報を生成し，後述する機械学習モデルに最初に学習させる．このツールは MAP-Elites による探索を行い，初期設定に基づいて DNA 構造体やその組み立て経路を多様に発見することができる．一定長の短い塩基配列を表す文字「ドメイン」が最小単位となっており，さらにそれらが 1 列に結合したものが「ストランド」である．ストランドの組み合わせ（ストランドセット）によっては互いに結合でき，多様な DNA 構造体をなす．

2.3 DA-QD

本研究では，MAP-Elites をベースに Lim ら [2] が提案した，「Dynamics-Aware Quality-Diversity（DA-QD）」(Fig.1) を参考にする．DA-QD ではより高速な「ダイナミクスモデルによる評価」を用いた仮想的な探索が先に行われ，「仮想スキルレパートリー」から実際に評価する対象が選ばれる．また，モデルは精度向上を狙いとして，ループのたびに新たに得られたデータを再学習する．本研究では DA-QD をベースとすることで，本来は全ストランドセットに対してシミュレーションを行い，その結果の評価を行ってレパートリー（QD のグリッド）への追加を行うところを，より高速に得られる「機械学習モデルによる予測結果」への評価によって先に仮想のレパートリーを作成し，予測の信頼度が低い場合に限ったシミュレーション適用が可能になる．

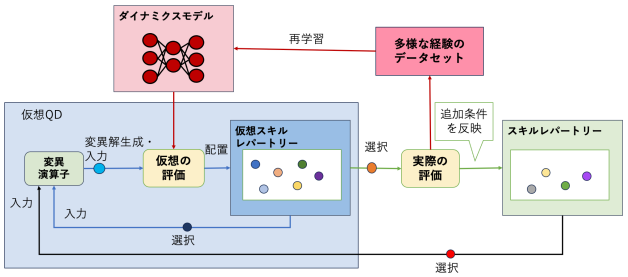

Figure 1: DA-QD の実行の流れ ([2] の Fig.1 を参考に作成)

3 モデル導入型 QD

本研究では，DA-QD をベースとした構造体の探索（モデル導入型 QD）が Fig.2 に示す流れで実行される．また，開発に必要な探索基準，使用シミュレータ，機械学習モデルは次のように決定した：

1) 探索基準は「構造体全体の平均位置エネルギー」（以後，単に位置エネルギーとする）という物理特性とし，その値が低くなるような構造体をなすストランドセットを探索する．位置エネルギーは低いほど構造体が安定しているとみなせる．

2) 位置エネルギーの計算にはシミュレーションツール「oxDNA」[3] を使用する．oxDNA は計算負荷が高いが，DA-QD を探索アルゴリズムのベースとすることで適用回数を削減し，探索アルゴリズム全体の効率化を行う．探索ツールから出力された構造体のドメインをランダムな塩基配列に置き換えることで oxDNA に入力でき，位置エネルギーが得られる．

3) 機械学習モデルとしては「サポートベクター回帰（SVR）」を導入し，ストランドセットからできうる全構造体の平均の位置エネルギーを予測する．この際，少しずつ異なるデータセットを学習した複数モデルの予測を統合する「バギング」によって，精度の向上を試みる．

モデル導入型 QD の流れは概ね DA-QD と共通で，SVR モデルによって位置エネルギーが予測され，QD グリッド

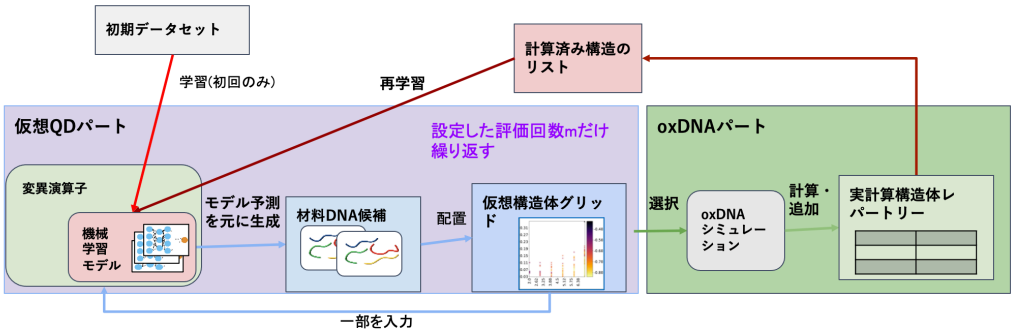

Figure 2: モデル導入型 QD の実行の流れ

が作成される．予測の信頼度が低ければ oxDNA によってエネルギーが計算され，モデルに再学習される．

4　実験の結果と考察

モデル導入型 QD と，oxDNA の計算だけでエネルギーを計算する「oxDNA 依存 QD」について，実行時間や oxDNA 実行回数を比較した．また，ループごとの用いた SVR モデルの精度の変化について考察した．

4.1　SVR モデルの精度変化

モデルによる予測精度は，3 回のループの間に表 1，2 のように変化した．RMSE，MAE，MAPE は正解と予測の誤差，R2 は相関の強さを表す．

Table 1: 単体の SVR の精度

実行回数	RMSE	MAE	MAPE	R2
ループ 0	0.03300	0.03947	3.03019	-6.67840
ループ 1	0.03197	0.03846	2.93408	-5.90491
ループ 2	0.03056	0.03725	2.80082	-3.95609

Table 2: バギングによる予測の精度

実行回数	RMSE	MAE	MAPE	R2
ループ 0	0.03462	0.04124	3.18564	-11.79754
ループ 1	0.03366	0.04016	3.09485	-9.90203
ループ 2	0.03154	0.03819	2.89562	-6.19785

単体のモデルとバギングのいずれにおいても，予測と正解の誤差は減少し，相関は増加する傾向にあり，再学習は有効である可能性がある．一方，R2 が負の値であることからもわかるように，この段階では予測精度は不足している上，バギングによりさらに精度が下がる点が不自然である．

4.2　oxDNA 依存 QD とモデル導入型 QD の実行時間とグリッド

モデル導入型 QD と oxDNA 依存 QD の実行速度と oxDNA 実行回数を表 3 に示す．

Table 3: モデル導入型 QD と oxDNA 依存 QD の効率性比較

アルゴリズム	モデル導入型 QD	oxDNA 依存 QD
初期ストランドセット数	318	317
実行時間 (s)	18131.89	250823.96
oxDNA 実行回数	90	2852
グリッドの解の数 (ループ 0,1,2)	39,48,43	31,25,27

また，モデル導入型 QD からは図 3 のようなグリッドが得られた．x 軸はストランド数，y 軸は 1 ストランドセットから得られる構造体のエネルギーにおける標準偏差（予測信頼度の低さ）である．

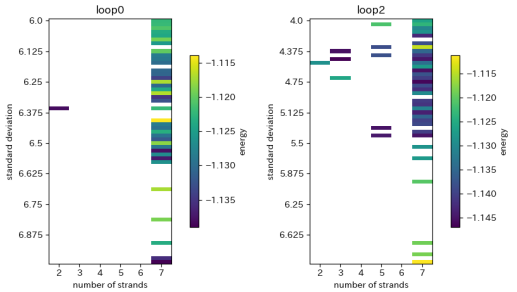

Figure 3: モデル導入型 QD の 1 回目の配置（左）と，3 回目の配置（右）により得られたグリッド

モデル導入型 QD の 3 ループの実行速度は，oxDNA 依存 QD のそれと比べるとはるかに高速である．また，ループを繰り返すごとにエネルギーの多様性が上がるように解が配置され，MAP-Elites が正常に機能している．一方，エリート解のストランド数は 7 に偏ってしまっている．

5　結論

DA-QD をベースに機械学習モデルを導入した QD はシミュレーションに依存した QD よりはるかに高速であり，再学習も有効である可能性がある．このことから，モデルの予測精度を確保できれば，DA-QD をベースとした探索により DNA 分子ロボットに使用できる塩基配列の探索が飛躍的に効率化されうると結論づけた．一方，モデルの精度不足やバギングによる精度低下，グリッドの解のストランド数の偏りなど，課題点も数多く残されている．バギングにおけるサンプリングやモデル選定，評価回数等に問題がないかを調査し，これらの原因を究明していく必要がある．

References

[1] L Cazenille, A Baccouche, and N Aubert-Kato. Automated exploration of dna-based structure self-assembly networks. *Royal Society Open Science*, Vol. 8, No. 10, p. 210848, 2021.

[2] Bryan Lim, Luca Grillotti, Lorenzo Bernasconi, and Antoine Cully. Dynamics-aware quality-diversity for efficient learning of skill repertoires. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5360–5366. IEEE, 2022.

[3] Petr Šulc, Flavio Romano, Thomas E. Ouldridge, Lorenzo Rovigatti, Jonathan P. K. Doye, and Ard A. Louis. Sequence-dependent thermodynamics of a coarse-grained dna model. *The Journal of Chemical Physics*, Vol. 137, No. 13, p. 135101, 2012.