Machine-learning-based prediction of DNA structure volume for Quality-Diversity exploration

百塚　真弥　(指導教員：オベル加藤ナタナエル)

1 Abstract

DNA nanotechnology enables the creation of nanoscale structures and devices using DNA's unique properties, including programmability, cost-effective synthesis, and nanoscale precision. A significant challenge remains in synthesizing diverse DNA strands to construct complex structures. This paper focuses on optimizing the size of DNA molecular assemblies as a proxy for molecular robot shapes, which significantly influence their movement. We develop a surrogate model seeded by physics-based simulations (oxDNA) to predict structure sizes, integrating this with a quality-diversity algorithm for iterative optimization. This process efficiently generates diverse candidate structures while minimizing computational costs. Additionally, we investigate the effects of temperature on DNA structures, exploring its impact on their properties and functionality.

2 Introduction

DNA nanotechnology leverages DNA's programmability, scalability, and cost-effective synthesis to create nanoscale structures and molecular devices, including molecular robots. These robots rely on components such as processors and actuators, using DNA as their core material. Practical applications range from drug delivery to environmental monitoring, emphasizing efficiency and adaptability. However, synthesizing unique DNA strands for diverse structures remains a significant hurdle. This thesis addresses the optimization of DNA molecular assemblies, focusing on size as a proxy for robot shapes. Using a quality-diversity approach, we integrate a surrogate model with physics-based simulations (oxDNA) to predict and refine structure sizes. Iterative optimization cycles improve the surrogate model and efficiently generate diverse candidates. Additionally, we examine how temperature influences DNA structures, aiming to understand their effects on their stability and functionality, which is critical for future molecular device applications.

3 Methods

We optimize DNA structures of various sizes using a quality diversity (QD) algorithm. Starting with basic DNA building blocks, we construct strand sets and use a surrogate model to predict structure sizes, avoiding direct evaluations with resource-intensive simulators like oxDNA. The surrogate model is initially seeded with datasets evaluated using oxDNA. The optimization alternates between QD-based exploration and surrogate model refinement, retraining with promising or mispredicted strand sets simulated by oxDNA. The workflow involves three stages Fig.1): 1. Generating an initial dataset. 2. Predicting DNA structure volumes and stability. 3. Creating diverse DNA strand sets based on predictions. This method efficiently predicts structure sizes and assembles DNA strands into various configurations.

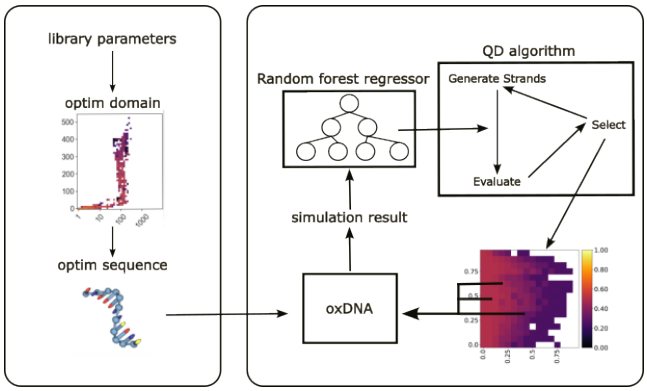

図 1: An overview of surrogate-based Quality Diversity algorithm

3.1 Generation of the preliminary dataset

We generate the initial dataset based on the previously mentioned library and optimize the domains using MAP-Elites [5] and Peppercorn [1]. Next, we select the top 10 domains with the highest reaction-type entropy (ERT) values and mean structural size (MSS). These domains are further optimized using NUPACK [3] and a genetic algorithm. As a result, we select a set of 20 DNA strands as the initial data for each library. This method is entirely based on [2].

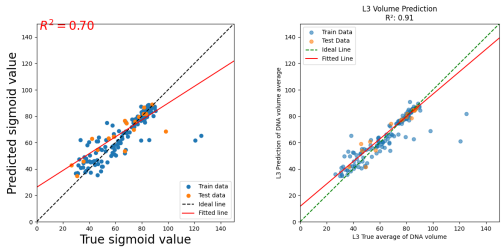

図 2: Prediction of volume and stability for L3

3.2 Surrogate Model: Prediction of DNA structure volume

In the Quality Diversity algorithm used in this study, the evaluation of individuals is based on the volume prediction of DNA structures, with the stability of the DNA structure as a feature. These values

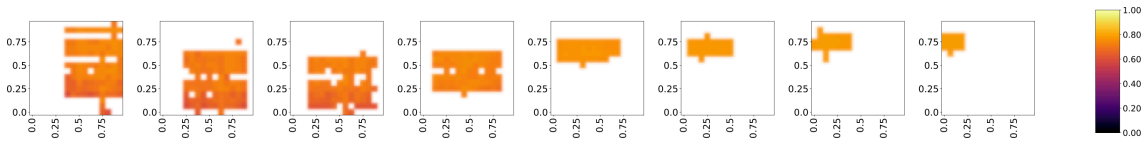

図 3: Results of Secondary QD Analysis

can be obtained using oxDNA, but running simulations for all individuals is computationally expensive. Therefore, we created a surrogate model to estimate the volume and stability values as a substitute.

3.3 Random Forest Regressor

We use the Random Forest Regressor for prediction. This regression method fits numerous decision trees to subsamples of the dataset and averages their outputs to control overfitting. Based on [4], regressors outperform neural networks for the current dataset. TensorFlow was used for implementation, with bootstrapping applied due to the small training dataset.

3.4 Inputs and Outputs

The model inputs consist of the initial dataset and DNA strand sets derived from the QD algorithm. These inputs include the type of DNA strand, temperature at the start of the simulation, and connectivity (second smallest eigenvalue of the graph). Outputs include the volume (calculated as the convex hull of nucleotide positions) and stability (based on binding transitions during simulations).

3.5 Mapping DNA strand sets

The MAP-Elites algorithm [5] was employed to generate diverse DNA strand sets producing DNA structures of varying volumes. Temperature was chosen as one of the features, as it significantly affects DNA structure size and behavior.

4 Results

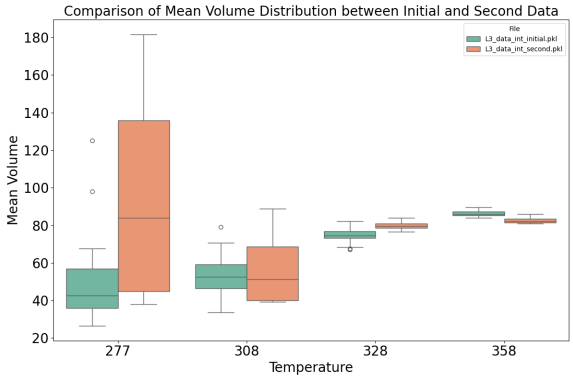

図 4: Comparison the volume of QD results and initial dataset

Fig. 2 presents volume predictions using the initial dataset for training and testing. R^2 values (0.7, 0.6, 0.8) indicate improved accuracy over the initial dataset, making the model suitable for further QD iterations.

Fig. 3 show DNA strand set generation, with each dot representing a set. The numbers 277, ..., 358 correspond to simulation temperatures. In each grid, the vertical axis is the free energy from NUPACK, and the horizontal axis is the predicted stability, both normalized to $[0, 1]$. The color indicates the predicted volume of DNA structures, also normalized.

As temperature increases, normalized free energy values decrease for all libraries, with $\frac{b}{c}$ becoming smaller. This suggests higher temperatures lead to increased free energy and decreased stability.

5 Conclusion and future work

This method allows for the efficient generation of DNA strand assemblies for DNA structures of various sizes. Additionally, the accuracy of volume and stability predictions for DNA structures in this workflow is considered sufficient. Moreover, obtaining DNA strand assemblies with specific characteristics, not just volume, is possible. Furthermore, using the DNA strand assemblies obtained from this workflow, I would like to investigate the relationship between the size of DNA structures and their interactions and further explore the optimal size of DNA structures for swarm behavior.

謝辞

本研究の一部は，JSPS 科研費 JP19KK0261 の助成を受けたものです．

参考文献

[1] Badelt, S., et al.: *J. R. Soc. Interface*, Vol. 17, No. 167, p. 20190866 (2020).

[2] Cazenille, L., Baccouche, A. and Aubert-Kato, N.: *J. R. Soc. Open Science*, Vol. 8, No. 10, p. 210848 (2021).

[3] Fornace, M. E., et al.: NUPACK: analysis and design of nucleic acid structures, devices, and systems (2022).

[4] Grinsztajn, L., Oyallon, E. and Varoquaux, G.: *Adv Neural Inf Process Syst*, Vol. 35, pp. 507–520 (2022).

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