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

Maya Hyakuzuka, Nathanael Aubert-Kato
Department of Information Sciences, Ochanomizu university, Japan
{g1920534, naubertkato}@is.ocha.ac.jp

Abstract

DNA nanotechnology has introduced the ability to create structures at the molecular scale, which is a promising approach for the implementation of very large swarms. However, the movement of such structures is heavily influenced by their size, prompting shape design optimization. Here, we use a quality-diversity approach to optimize the size of structures assembled from sets of DNA strands. We introduced a surrogate model to accelerate evaluations, with the ground truth provided by oxDNA, a physics-based simulator. We then iterate between optimization rounds using the QD algorithm, direct evaluation of promising and potentially mispredicted sets with oxDNA, and training of the surrogate model. We show that this approach efficiently generates diverse candidate sets at a fraction of simulation costs. Additionally, the surrogate model is reusable, enhancing the overall performance of future optimization tasks.

Introduction

DNA nanotechnology is a rapidly growing field that uses DNA molecules to create structures and dynamic devices at the nanoscale (Seeman and Sleiman, 2017). Specifically, the sequence of DNA molecules dictate how they interact with each other. Moreover, single-stranded DNA is flexible, while double-stranded DNA is more rigid, allowing the creation of dynamic devices (DeLuca et al., 2020). Besides applications to the biological and medical fields, those properties make DNA structures a tantalizing substrate for the creation and evaluation of large swarms of agents, going far beyond the numbers possible with electronic devices.

DNA nanostructures are an important aspect of the field of molecular robotics, offering the potential for smart task-specific systems (Nummelin et al., 2020). Swarming behaviors, a typical application for the large number of units available at the molecular scale, are heavily influenced by the size and shape of these nanostructures, as highlighted by (Kabir et al., 2020). As such, it is important for practical application to generate families of structures at a variety of sizes, depending on the task.

However, designing specific structures is not simple. Any part of a given DNA strand can potentially interact with any other element of the system, prompting careful consideration to avoid unwanted structures. The usual approach is to use specific building blocks to reduce the design space, combined with computer-assisted design methods (Rothemund, 2006; Ong et al., 2017). However, those rational approaches tend to limit the range of created structures to specific targets, therefore reducing the potential for the emergence of complex behaviors in the system. Here, we instead extend a recent approach by Cazenille *et al.* that relies on a Quality-Diversity algorithm to find sets of DNA strands that would assemble into rich families of structures, rather than a specific one (Cazenille et al., 2021).

The main limitation of the approach of Cazenille *et al.* is the lack of validation of the actual size of the structure generated, relying instead on a theoretical graph of bonds. Direct validation requires either computationally expensive simulations or time-consuming experiments. Here, we propose to overcome that limitation through a surrogate model

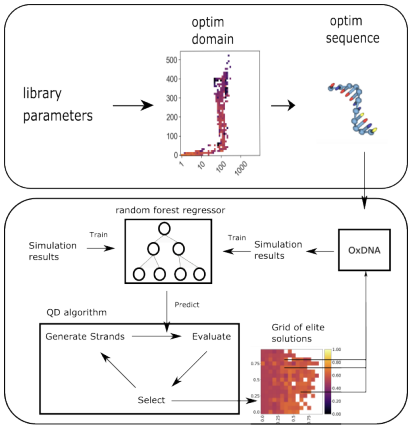
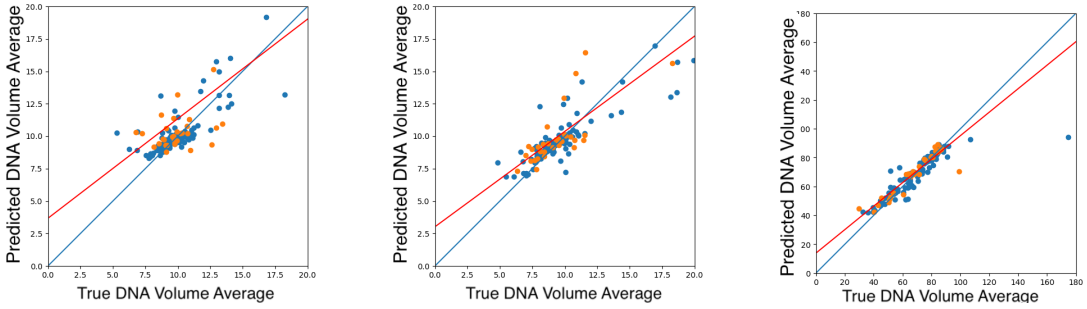


Figure 1: Workflow of our approach: we use a pre-established initial dataset of DNA sequences (Cazenille et al. (2021); top box) to train a surrogate model that avoids costly simulations of DNA structure dynamics. That model is then used for the optimization of sets of DNA strands. Sets that may have been wrongly evaluated or seem promising are eventually simulated and used to further train the model.



(a) L1 (b) L2 (c) L3

Figure 2: Surrogate performance evaluation for L1, L2, and L3. The x axis represent the evaluation obtained through OxDNA ("true" evaluation) and the y axis the prediction from our final model for training data (orange; taken both from the initial strands sets and the iterations of the algorithm) and test data (blue; selected randomly from the same data set). The red line shows the least-square fit of the data. The $x = y$ line (perfect match) is shown for convenience in blue.

for the fast prediction of structure sizes. Our workflow has three steps: (1) train a surrogate model on an initial dataset by evaluating promising sets of strands previously identified by Cazenille *et al.* with the OxDNA simulator (Ouldridge et al., 2011), (2) perform a Quality-Diversity optimization run, using that surrogate model for evaluation, (3) automatically select sets that are either promising or suspected to be incorrectly predicted, evaluate them with the simulator, and use those additional data to further train the model (Figure Figure 1).

Our approach was inspired by the Dynamics-Aware Quality-Diversity (DA-QD) algorithm (Lim et al., 2022), which progressively trains a model of the behavior of the optimized system to reduce the number of costly evaluations. We found that our approach led to the creation of an accurate model for the sets evaluated and an efficient exploration of the design space.

Method

Strands sets

Sets are created by selecting 2 to 7 strands from the three libraries defined by Cazenille *et al.*, denoted as L1 (short, simple strands), L2 (short strands with more variations), and L3 (medium length strands with large variations).

DNA structures volume evaluation

We use OxDNA, a coarse-grained simulator for DNA molecular dynamics (Ouldridge et al., 2011), to evaluate the size of structures created by a given set of strands. 5 copies of each strands are mixed together and simulated for 200000 time steps. We then take the average volume of the convex hull of all individual structures in the environment.

DNA structures volume prediction

We use a random forest regressor to predict the volume of DNA structures generated by a specific set. Features are the strands present in the set (encoded as a binary string), the temperature of the system, and the connectivity (Anderson Jr and Morley, 1985) of the graph of bonding reactions between strands. We chose that regressor as it was shown to be one of them most accurate on such data type (Grinsztajn et al., 2022).

DNA strands sets generation

We use MAP-Elites (Mouret and Clune, 2015), a Quality-Diversity algorithm, to find sets of strands that are high performing (generate large structures) with respect to specific features (diversity characteristics, also used as inputs to the prediction model), namely the total free energy in the system (reactivity) and the ratio of GC content (Guanine or Cytosine, corresponding to strong binding).

We use three mutation operators which repectively add a strand, remove a strand, or change the temperature of the system (bonding dynamics).

Results

Prediction results for the surrogate model after training through an optimization run are shown in Figure 2. We found generally a good agreement between the predicted volume and evaluated volume from OxDNA. The exploration could find a wide range of high performing sets, filling up most of the search grids. A typical result is shown at the bottom of Figure 1.

Conclusion

We introduced a surrogate model to implement a Quality-Diversity exploration of nanostructures made of DNA. Those structures have the potential to implement swarms of millions of individuals at the nanoscale. The main limitation of our current approach is the few number of data points available (128 training, 32 testing).

Data Availability

Code and data are available at `https://doi.org/10.5281/zenodo.11467097`.

References

Anderson Jr, W. N. and Morley, T. D. (1985). Eigenvalues of the laplacian of a graph. *Linear and multilinear algebra*, 18(2):141–145.

Cazenille, L., Baccouche, A., and Aubert-Kato, N. (2021). Automated exploration of dna-based structure self-assembly networks. *Royal Society Open Science*, 8(10):210848.

DeLuca, M., Shi, Z., Castro, C. E., and Arya, G. (2020). Dynamic dna nanotechnology: toward functional nanoscale devices. *Nanoscale Horizons*, 5(2):182–201.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Kabir, A. M. R., Inoue, D., and Kakugo, A. (2020). Molecular swarm robots: recent progress and future challenges. *Science and technology of advanced materials*, 21(1):323–332.

Lim, B., Grillotti, L., Bernasconi, L., and Cully, A. (2022). Dynamics-aware quality-diversity for efficient learning of skill repertoires. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5360–5366. IEEE.

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

Nummelin, S., Shen, B., Piskunen, P., Liu, Q., Kostiainen, M. A., and Linko, V. (2020). Robotic dna nanostructures. *ACS Synthetic Biology*, 9(8):1923–1940.

Ong, L. L., Hanikel, N., Yaghi, O. K., Grun, C., Strauss, M. T., Bron, P., Lai-Kee-Him, J., Schueder, F., Wang, B., Wang, P., et al. (2017). Programmable self-assembly of three-dimensional nanostructures from 10,000 unique components. *Nature*, 552(7683):72–77.

Ouldridge, T. E., Louis, A. A., and Doye, J. P. (2011). Structural, mechanical, and thermodynamic properties of a coarse-grained dna model. *The Journal of chemical physics*, 134(8).

Rothemund, P. W. (2006). Folding dna to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302.

Seeman, N. C. and Sleiman, H. F. (2017). Dna nanotechnology. *Nature Reviews Materials*, 3(1):1–23.