Third Year Research Internship (2024 - 2025)

Impact of the geometry of 3D molecular structures of synthetic DNA strands in the context of DNA data storage

Keywords: DNA Storage – 3D Point Clouds – Geometrical Similarity – AI

Main Skills: Good knowledge of programming (Python) and algorithmic geometry. Knowledge of machine learning and AI would be an asset.

Location: i3S laboratory (CNRS - Université Côte d'Azur) - MediaCoding Team - Sophia Antipolis, France.
Start of the internship: March 2025 - **Period**: 5 – 6 months.

Funding: Gratification based on official rate (~ 650€ per month)
PhD continuation: possible

Supervision / Contact:
- Frédéric Payan (frederic.payan@univ-cotedazur.fr)
- Marc Antonini (marc.antonini@cnrs.fr).

Internship Context

To store so-called 'cold' digital data, the limited lifespan of current storage media requires eco-responsible solutions. One such solution is storage on synthetic DNA [DA22], which transforms binary digital data into a sequence of quaternary symbols (noted as A, T, C, and G, referencing nucleotides). This sequence is then synthesized into DNA strands, ultimately stored in a sealed capsule resistant to environmental conditions.

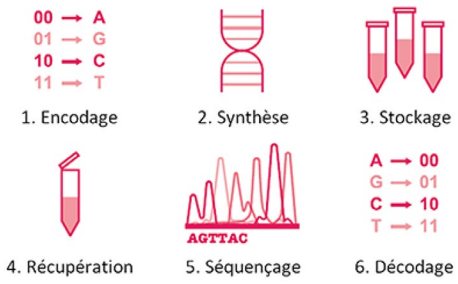

Fig. 1: Synthetic DNA storage process. Image from [PEPR].

A current limitation of this solution is the potential introduction of errors during the retrieval of the initial digital data. These errors may arise from incorrect reading of the quaternary symbols during DNA sequencing, a complex biochemical process needed before decoding the quaternary, then binary, symbols.

Numerous studies are currently focused on eliminating errors introduced during the sequencing of synthetic DNA strands. One of the most ambitious projects in this field is the ongoing PEPR MoleculArXiv [PEPR], led by the I3S laboratory since 2022, which will fund the proposed internship.

Internship Topic

DNA storage can be viewed as transmission over a noisy channel. The solutions proposed so far to reduce and/or correct DNA storage errors [Dim20] are logically inspired by the popular error-correcting codes based on information theory and widely used in digital transmission.

In recent years, I3S has collaborated on the SENSAAS project, which, drawing inspiration from computer vision and graphics techniques, has led to the development of the eponymous software [SensaasGit]. This software superimposes molecular structures by detecting their 3D geometric similarities [DP22]. This technique is based on the principle that "*the 3D geometric shape of a molecule plays a crucial role in its interaction with the environment*" [KKE09].

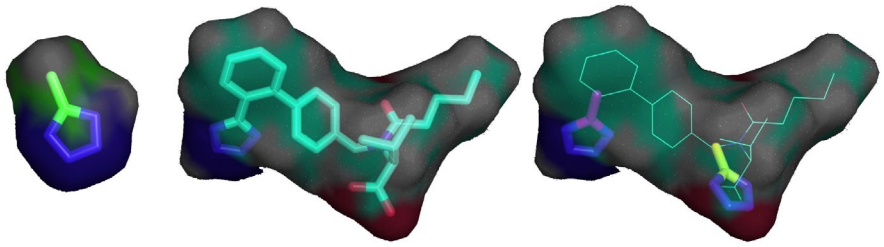

Fig. 2 : Example of molecular structure superimposition obtained with Sensaas by registration of 3D point clouds.

The goal of this internship is to address the following question: In the context of synthetic DNA storage, does the 3D geometric shape of molecular structures constituting DNA strands play a role in the occurrence of errors during sequencing/decoding?

To attempt to answer this question, the intern will use the SENSAAS software [SensaasGit] on 3D molecular structures (point clouds) representing DNA strands. He/she may also utilize recently developed AI solutions for 3D point clouds [ZWT+23].

References

[PEPR] *Programmes et équipements prioritaires de recherche* (PEPR) MoleculArXiv. https://pepr-molecularxiv.fr, France 2030.

[DA22] M. Dimopoulou et M. Antonini. *Data and image storage on synthetic DNA: existing solutions and challenges*. In EURASIP J. Image Video Process., oct. 2022.

[DP22] D. Douguet, F. Payan, *Sensaas: Shape-based Alignment by Registration of Colored Point-based Surfaces*. In Mol. Inf., jun. 2020. https://doi.org/10.1002/minf.202000081.

[Dim20] M. Dimopoulou. *Encoding techniques for long-term storage of digital images into synthetic DNA*. Doctorat de l'Université Côte d'Azur, dec. 2020.

[KKE09] S. Kortagere, M. D. Krasowski et S. Ekins. *The importance of discerning shape in molecular pharmacology.* In Trends Pharmacol., 2009.

[SensaasGit] Github SenSaaS: https://github.com/SENSAAS/sensaas.

[ZWT+23] H. Zhang, C. Wang, S. Tian *et al.*, *Deep learning-based 3D point cloud classification: A systematic survey and outlook*. In Displays, sep. 2023. https://doi.org/10.1016/j.displa.2023.102456.

Stage de Recherche 3ème année (2024 - 2025)

Impact de la géométrie de la structure moléculaire 3D d'un brin d'ADN synthétique dans le contexte du stockage de données sur ADN

Mots - clefs : Stockage sur ADN – Nuages de Points 3D – Similarité Géométrique – IA

Compétences recherchées : Bonnes connaissances en programmation (Python) et géométrie algorithmique. Des connaissances en Machine Learning et IA seraient des atouts.

Lieu : Laboratoire i3S (CNRS - Université Côte d'Azur) – Equipe MediaCoding - Sophia Antipolis, France.
Début du stage : Mars 2025 - **Période** : 5 – 6 mois

Gratification : fixé sur le barème officiel (~ 650€ per month)
Possibilité de poursuite en thèse : possible
Encadrants :

- Frédéric Payan (frederic.payan@univ-cotedazur.fr)
- Marc Antonini (marc.antonini@cnrs.fr).

Contexte du stage

Pour stocker des données numériques dites "froides", la durée de vie limitée des supports actuels exige que l'on propose des solutions éco-responsables. L'une d'entre elles est le stockage sur ADN synthétique [DA22], dont le principe est de transformer une donnée numérique en une séquence de symboles quaternaires (notés A, T, C et G en référence aux nucléotides) pour ensuite la synthétiser en brins d'ADN stockés dans une capsule étanche et résistante aux conditions environnementales.

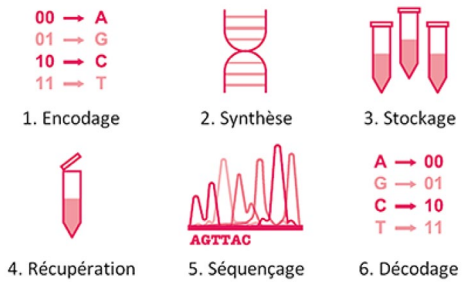

Figure 1 : Processus de stockage sur ADN synthétique. Image issue de [PEPR]

Une limite actuelle de cette solution est l'introduction potentielle d'erreurs au moment du désarchivage de la donnée numérique initiale. Ces erreurs peuvent provenir d'une lecture erronée des symboles quaternaires lors du séquençage de l'ADN, qui est un processus biochimique complexe, nécessaire avant le décodage des symboles quaternaires, puis binaires.

De nombreuses recherches sont menées actuellement pour supprimer les erreurs introduites lors du séquençage des brins d'ADN synthétique. L'un des projets les plus ambitieux dans ce domaine est l'actuel PEPR MoleculArXiv [PEPR] piloté par le laboratoire I3S depuis 2022, et qui financera le stage proposé présentement.

Sujet du stage

Le stockage sur ADN peut être vu comme une transmission sur un canal bruité. Les solutions proposées jusqu'à maintenant pour réduire et/ou corriger les erreurs de stockage sur ADN [Dim20] sont logiquement inspirés des populaires codes correcteurs d'erreurs largement utilisés en transmission numérique qui s'appuient sur la théorie de l'information.

Ces dernières années, I3S a collaboré au projet SENSAAS qui, en s'inspirant de techniques de vision par ordinateur et d'informatique graphique, a donné naissance au logiciel éponyme [SensaasGit]. Ce logiciel permet d'aligner des structures moléculaires en détectant leurs similarités géométriques 3D [DP22]. Cette technique part du principe que « *la forme géométrique 3D d'une molécule joue un rôle primordial dans l'interaction avec son environnement* » [KKE09].

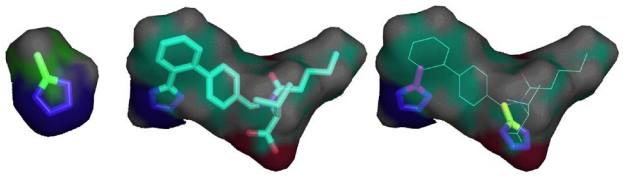

Figure 2 : Exemple de superposition de structures moléculaires obtenue avec Sensaas par recalage de nuages de points 3D.

L'objectif de ce stage sera de répondre à la question suivante : dans le contexte du stockage sur ADN, est-ce que la forme géométrique 3D des structures moléculaires constituant les brins d'ADN joue un rôle dans l'apparition d'erreurs lors du séquençage /décodage ?

Pour tenter de répondre à cette question, le/la stagiaire utilisera le logiciel SENSAAS [SensaasGit] sur des structures moléculaires 3D (nuages de points) représentant des brins d'ADN. Il/Elle pourrait également être amené(e) à utiliser des solutions d'IA récemment développées pour les nuages de points 3D [ZWT+23].

Références

[PEPR] *Programmes et équipements prioritaires de recherche* (PEPR) MoleculArXiv. https://pepr-molecularxiv.fr, France 2030.

[DA22] M. Dimopoulou et M. Antonini. *Data and image storage on synthetic DNA: existing solutions and challenges*. In EURASIP J. Image Video Process., oct. 2022.

[DP22] D. Douguet, F. Payan, *Sensaas: Shape-based Alignment by Registration of Colored Point-based Surfaces*. In Mol. Inf., jun. 2020. https://doi.org/10.1002/minf.202000081.

[Dim20] M. Dimopoulou. *Encoding techniques for long-term storage of digital images into synthetic DNA*. Doctorat de l'Université Côte d'Azur, dec. 2020.

[KKE09] S. Kortagere, M. D. Krasowski et S. Ekins. *The importance of discerning shape in molecular pharmacology*. In Trends Pharmacol., 2009.

[SensaasGit] Github SenSaaS : https://github.com/SENSAAS/sensaas.

[ZWT+23] H. Zhang, C. Wang, S. Tian *et al.*, *Deep learning-based 3D point cloud classification: A systematic survey and outlook*. In Displays, sep. 2023. https://doi.org/10.1016/j.displa.2023.102456.