DONE: **D**ual h**O**t e**N**coding for biological sequences

During last summer internship 2024, we introduced a novel encoding scheme for omics sequences, integrating Komlos and Hadamard transforms. Unlike traditional one-hot encoding, this approach offers a more informative representation of omics data while significantly reducing computational complexity. By leveraging the inherent properties of these transforms, our method effectively captures complex patterns within the data, leading to improved model accuracy and reduced training times. When combined with a Hilbert Curve-based spatial mapping, this encoding scheme demonstrates particularly efficient results, achieving superior performance across various predictive tasks with significantly lower computational resource demands compared to one-hot encoding. This translates to faster training times and reduced reliance on CPU and GPU resources. Our findings suggest that this novel encoding scheme, particularly when integrated with Hilbert Curve mapping, holds significant promise for advancing omics data analysis by offering a more efficient and effective approach to feature representation. The outcome of the 2024 summer internship was submitted to ISMB/ECCB.

The focus of this summer internship is to extend this model to encode the proteins or peptide sequences. We build the foundation of the algorithm and we wanted to write the code in Python or R and apply it for proteins solubility or protein crystallization.

Project Type: Research

Internship Batch:

- **Batch 1:** May 11 to July 10, suitable for Education City students, i.e., CMUQ, TAMUQ and HBKU students

Expected Learning Outcomes

- Exposure to DNA and proteins sequence data.
- Hands-on experience with AI and ML.
- Training in coding (Python/R), machine learning

Preferred Skills

- Coding in R or Python. Help in writing of the result section for the manuscript to be submitted.

Mentors

Name: Dr. Halima Bensmail email: hbensmail@hbku.edu.qa