ENGINEERING DATA STRUCTURE ORGANOIDS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 899833

PORTABOLOMICS

Portabolomics Strand:
1.1, 1.2, 2.1

J. Connolly, B. Shirt-Ediss, E. Torelli, J. Bacardit, N. Krasnogor

X X

Artificial Intelligence techniques for Nucleic Acid Origami experiments

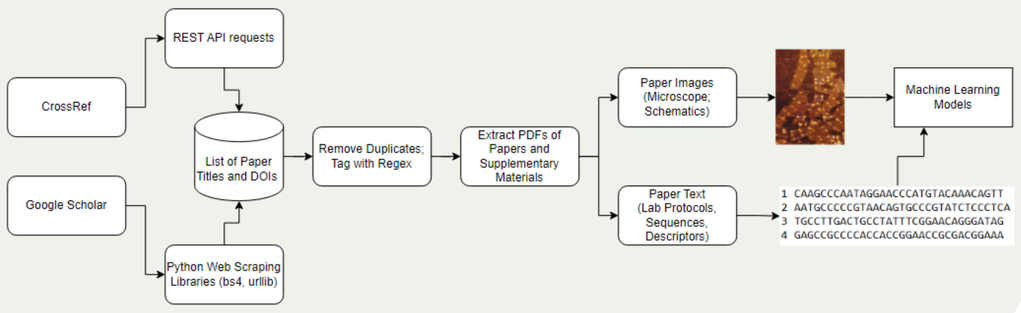
Abstract:

Potential applications are increasingly shown in the literature enabled by the advent of scaffolded DNA origami [1]. We employ a data-driven approach, applying machine learning to a curated a database collected from Nucleic Acid Origami literature. We expect to aid the design of well-formed nucleic acid origami through machine learning informed lab protocols and algorithm improved sequence design. This will lead to improved application of nucleic acid origami, with increased yield, scale and complexity.

Aim of the project:

To develop strategies driven by artificial intelligence techniques, to assist the design of Nucleic Acid Origami sequences and experimental protocols.

1) Create a curated corpus of literature data.
2) Create Automated Quantification of well-formed Nucleic Acid Origami.
3) Determine causes of well-formed Origami.
4) Use generative algorithms to identify and overcome design bottlenecks.



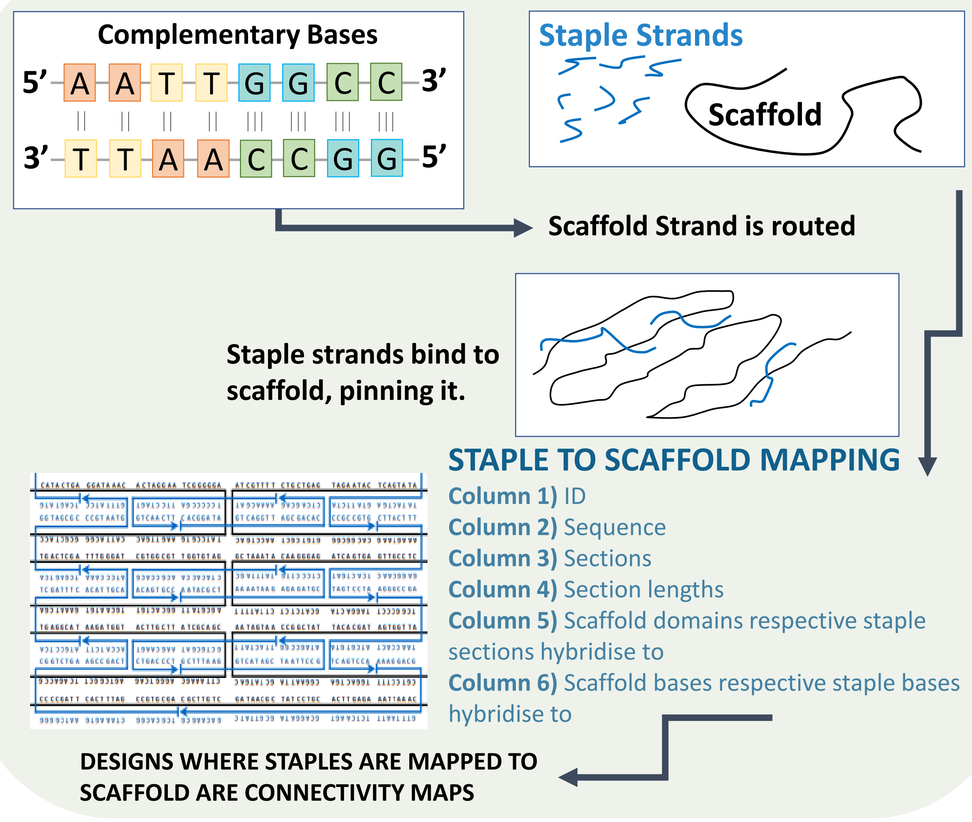
Achievements throughout the project:

- A database of Nucleic Acid Origami experiment data was manually curated from 552 relevant research papers producing over 1115 unique instances for use in downstream computational tasks.

-This allowed us to produce a minimal information representation of a nucleic acid origami experiment (NAOMEE: Nucleic Acid Origami Minimal Exchange Format).

-We performed systematic machine learning experiments to find the best pre-processing pipeline [2] and models for prediction of important lab protocols.

-We tackled the generation of optimal scaffold and staple sequence strands for production of higher yield Nucleic Acid Origami using a multi-objective optimisation algorithm using negative heuristic metrics.



Outcomes:

- Co-authored conference poster + technical paper (DNA 26, [3])
- First Author of conference poster for (UCNC 2021)
- Co-authored a conference poster (DNA 28)
- First Author of a conference poster + satellite talk (ICSB 2022)
- First Author of a conference poster (SBUK 2022)
- Produced a curated literature data base of experimental data.

	r2	MAE	MSE	RMSE
Magnesium (mM)	0.636 ± 0.013	1.105 ± 0.028	5.224 ± 0.184	2.280 ± 0.039

	r2	MAE	MSE	RMSE
Staple Molarity (nM)	0.435 ± 0.093	44.477 ± 5.325	24249.345 ± 3920.128	152.81 ± 13.923

	Accuracy	Recall	Precision	F1 Score	ROC AUC
Thermal Profile (Binary)	0.617 ± 0.009	0.616 ± 0.016	0.482 ± 0.014	0.535 ± 0.013	0.608 ± 0.009

Predictive models were created using the Extra Trees Algorithm with a 3x and 5x Cross-Validation (n=30 repetitions). Results shown are full data-set (550 papers) 3x cross-validation using appropriate Pre-processing and Recursive Feature Elimination pipelines.

Open questions and future directions:

- Work has started on a position paper, based upon the curated data collected, to give direction towards correct and reproducible representations of nucleic acid origami data.
- Using our design and image data, we have planned a computational pipeline for creating synthetic images to feed an automated quantification model, leading to improved Nucleic Acid Origami metrics.

References:

[1] https://www.nature.com/articles/nature04586 [2] https://www.nature.com/articles/s42256-019-0138-9 [3] RevNano DNA26 conference (http://dna26.iopconfs.org/posters)

This work is funded by EPSRC Grant EP/N031962/1. Synthetic Portabolomics: Leading the way at the crossroads of the Digital and the Bio Economies

EPSRC Pioneering research and skills

Centre For Bacterial Cell Biology

I(C)2S

Policy Ethics And Life Sciences

Newcastle University