SPMS20014 – Predicting G-quadruplex Formation *in vivo* and *in vitro* with Deep Convolutional Neural Network

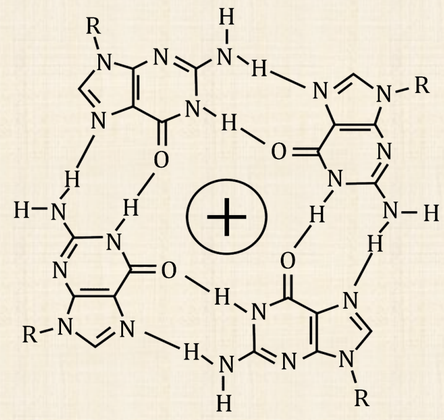
Presented by Mr Oon Yu Yang

Supervised by Prof Phan Anh Tuân, Ms Anna Korsakova

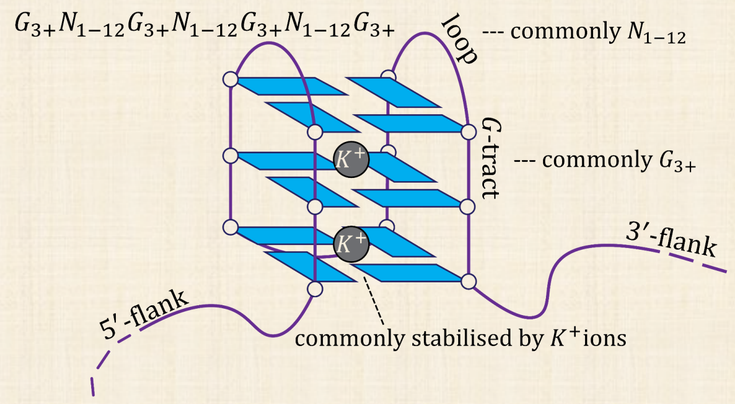
Introduction

G-quadruplexes (G4s) are a class of stable DNA and RNA secondary structures formed in guanine-rich nucleic acid sequences which consist of guanine tracts (G-tracts) with connecting loops and are stabilised through Hoogsteen hydrogen bonding within and π-π stacking between quartets of guanines (G-tetrads).

Due to their importance as mediators of genomic functions, G4s have been studied extensively, and numerous computational approaches have been developed for their identification. However, many widely accepted putative quadruplex sequences (PQS) *in vitro* do not actually form stable genomic G4s *in vivo*.



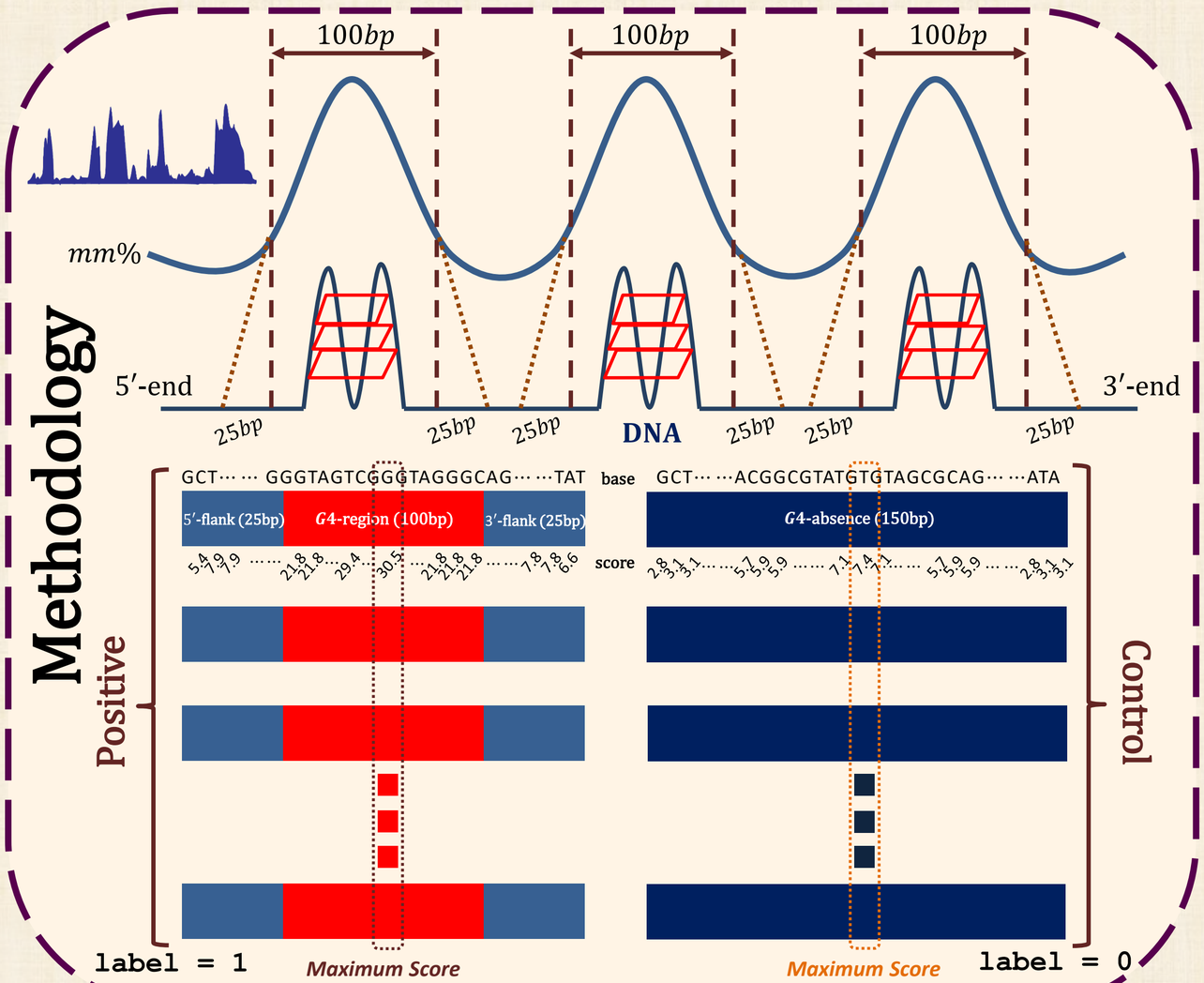
A structural representation of a planar G-tetrad.



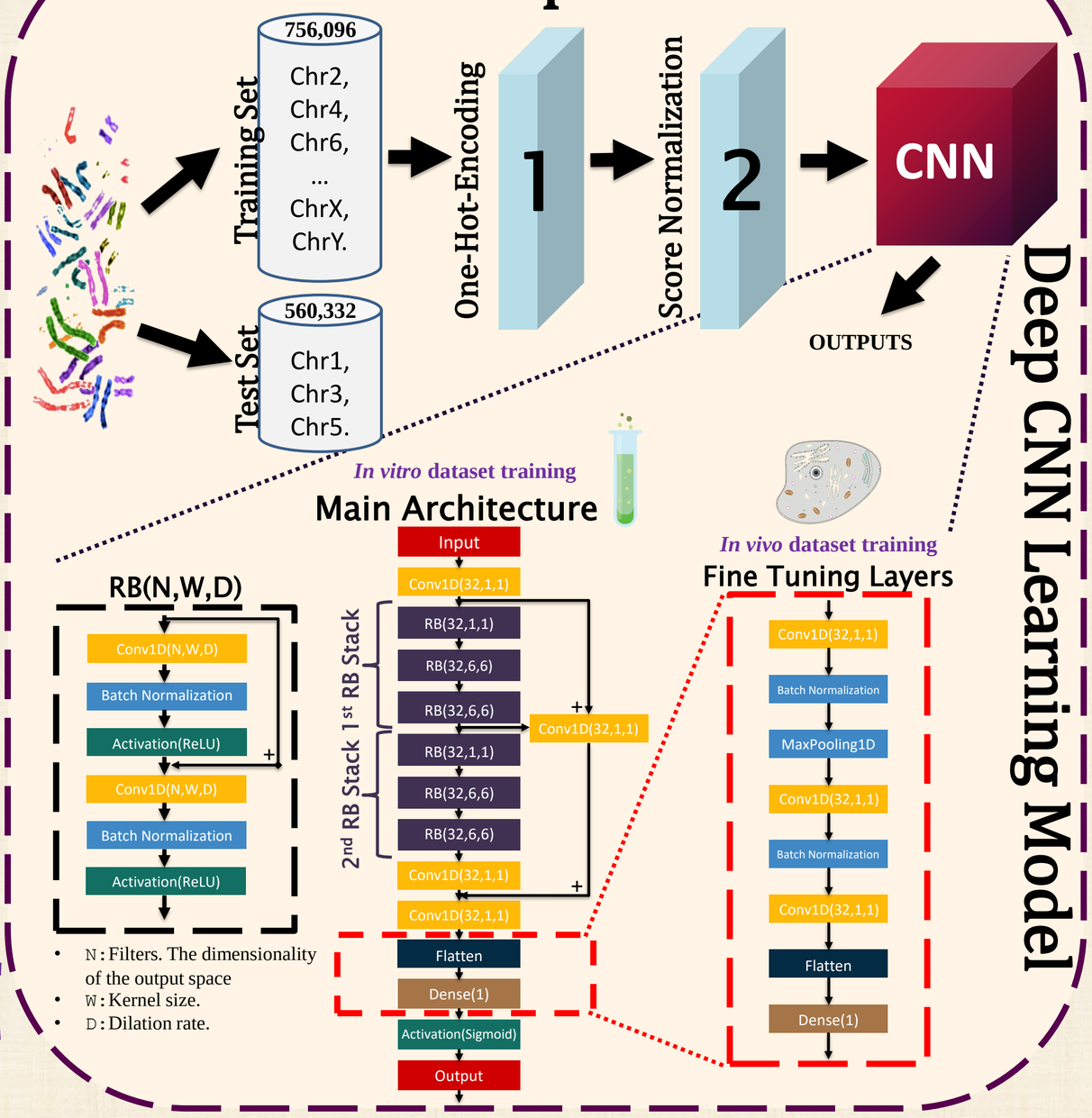
A schematic representation of a single strand DNA sequence forming a G-quadruplex.

Here, we present a deep learning method based on multi-layered convolutional neural networks (CNN) that can learn the characteristics of G4 sequences and predicts the formation of G4s with high propensity which outperforms state-of-the-art models.

Methodology

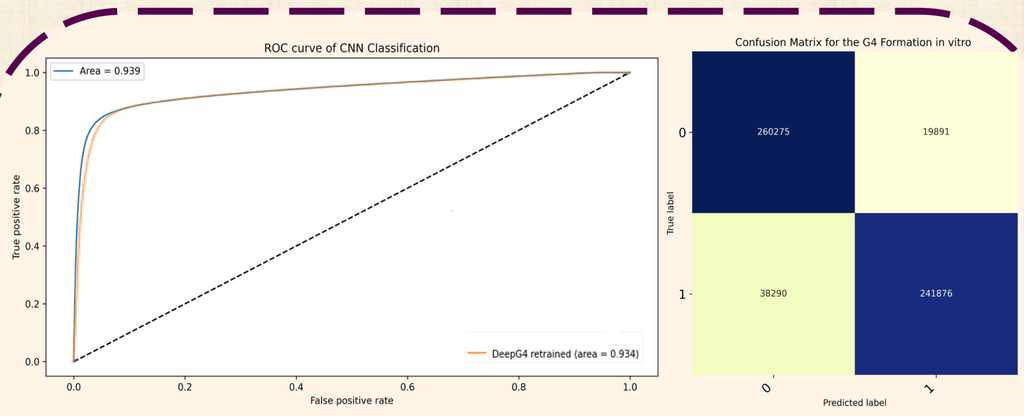


Data Preparation



- N : Filters. The dimensionality of the output space
- W : Kernel size.
- D : Dilation rate.

Results



(Left) Predictions of the *in vivo* formation of the G4s using DeepG4 model from Rocher et. al. (2020) (orange curve) and our model, which was pretrained on *in vitro* dataset and fine-tuned on the *in vivo* dataset (blue curve). Although both models were trained on the same dataset, the inputs and labels preparation methods were different. **(Right)** The confusion matrix showing the performance of our Deep CNN model on the test dataset, with accuracy of 89.62% and F1-score 0.893.

Acknowledgement

The author would like to express his most sincere gratitude and appreciation to his supervisors, Prof Phan Anh Tuân and Ms Anna Korsakova, for their continuous support and encouragement, without whom, this project would not have been possible.

Future Works

Although RNA G-quadruplexes (rG4s) share the main structural characteristics with DNA G-quadruplexes, study reveals that rG4s possess higher thermodynamic stability than their DNA counterparts. Future research shall focus on extending the fine-tuning approach to rG4s dataset published by Kwok et al. (2016), where data are scarcer, and therefore, transfer learning will be of value.

Citations

- Daniela Rhodes and Hans J. Lipps, Nucleic Acids Research **43** (18), 8627 (2015).
- Dipankar Sen and Walter Gilbert, Nature **334** (6180), 364 (1988).
- J. L. Huppert, Nucleic Acids Research **33** (9), 2908 (2005).
- Aleksandr B. Sahakyan, Vicki S. Chambers, Giovanni Marsico, Tobias Santner, Marco Di Antonio, and Shankar Balasubramanian, Sci Rep **7** (1), 14535 (2017).
- Chun Kit Kwok, Giovanni Marsico, Aleksandr B. Sahakyan, Vicki S. Chambers, and Shankar Balasubramanian, Nat Methods **13** (10), 841 (2016).
- Vincent Rocher, Matthieu Genais, Elissar Nassereddine, and Raphael Mourad, preprint, 2020.