|  |
| --- |
| **Research Theme: Computational Biology; AI; MIcrobiology** |
| **PhD Research Project Title:** **Predicting DNA-Protein Binding via AI to Decode Gene Regulatory Networks** |
| **Scholarship category (Please indicate the source of funding for this project):**1. **SBS Research Student Scholarship (for SBS faculty only)**
 |
| **Principal Investigator/Supervisor: Anni Zhang** |
| **Co-supervisor/ Collaborator(s) (if any):**  |
| **Project Description****a) Background:** Interpretation of gene regulatory networks enables the prediction of cellular behavior in response to environmental factors and interactions with other organisms. Predicting DNA-protein binding, especially identifying the binding sites of transcription factors (TFs), is an effective method to decode gene regulatory networks from genomes. However, experimentally identifying the binding sites for a TF often requires extensive laboratory work. For example, to interpret regulatory networks in *Pseudomonas aeruginosa*, one study combined chromatin immunoprecipitation sequencing (ChIP-seq) with RNA-seq on *P. aeruginosa* mutants. For each TF, a mutant was created to overexpress the TF, resulting in regulatory profiles for only 20 TFs (Huang *et al.* 2019). In addition, most studies focus on model organisms, as ChIP-seq requires at least 1 million cells, a quantity often unavailable from not-yet-cultured species.Computational methods, such as motif identification, comparative genomics, and machine learning models, offer an alternative for predicting TF binding sites. However, they often require extra information about the TF.More importantly, most methods analyzed each binding site nucleotide base separately. However, binding involves the interaction among multiple amino acids and nucleotide bases. Without considering the spatial arrangement between protein and DNA, this approach limits the generalization of DNA-protein binding patterns for TFs without prior knowledge.In summary, neither experimental nor existing computational methods can efficiently predict DNA-protein binding. Thus, there exists a *critical need* to design novel methods that 1) consider interactions between protein domains and DNA binding sites for more precise prediction, and 2) effectively generalize DNA-protein binding patterns for under-characterized TFs. **b) Proposed work:**Our *overall objective* is to design AI models to predict DNA-protein binding patterns for diverse TF-binding site pairs without the requirement of prior knowledge. Our *central hypothesis* is that AI models, such as language models, allow accurate prediction of DNA-protein binding. Our central hypothesis is based upon previous work and our preliminary data:1. A language model BERT trained on DNA sequences demonstrated high accuracy (AUC = 94.7%) in predicting DNA-protein binding within human genomes (Luo *et al.* 2023).
2. A transformer-based capsule network reported high accuracy (AUC > 91%) in predicting TF binding sites for all five human cell lines (Ghosh *et al.* 2023).
3. Deep neural network models have shown high accuracy in predicting binding site sequences for *Saccharomyces cerevisiae* (Pearson’s r = 0.960, P < 5 × 10−324, n = 61,150), facilitating expression engineering (Vaishnav *et al.* 2022).
4. Our *prototype models*, trained on *E. coli* DNA-TF binding pairs, achieved a prediction accuracy of 73.7%. The model successfully predicted higher binding probabilities for true positives than for true negatives in two of the three TF families (compared to one by AlphaFold3). Additionally, it identified the complex DNA-TF dimer binding structure and accurately recognized the known DNA binding domain (Fisher's test, p = 0.001).

**c) Preferred skills:** **A curiosity about microbes;** **An interest in coding;** **A dedication to research** |
| **Supervisor contact:****If you have questions regarding this project, please email the Principal Investigator: anni.zhang@ntu.edu.sg** |
| **SBS contact and how to apply:**Associate Chair-Biological Sciences (Graduate Studies) : AC-SBS-GS@ntu.edu.sg Please apply at the following: **Application portal:** <https://venus.wis.ntu.edu.sg/GOAL/OnlineApplicationModule/frmOnlineApplication.ASPX> |