IFN712 Research Project Form

(Submitted to [y.feng@qut.edu.au](mailto:y.feng@qut.edu.au) by 30 June 2025)

|  |  |
| --- | --- |
| Project agency (School, industry, funded/HDR) | QUT School of Computer Science |
| Industry/project supervisor and contact emails |  |
| Academic Supervisor name(s) and contact emails | Dr Jake Bradford  [jake.bradford@qut.edu.au](mailto:jake.bradford@qut.edu.au) |
| Information Technology major(s) | Artificial Intelligence  Computer Science  Data Science  Software Development |
| Project title | The Power of Patterns: Using Binary Codes to Improve the Safety of the CRISPR Gene Editing Technology |
| Brief description of the research problem, aims, method and expected outputs (100~200 words) | In nature, the CRISPR system functions as part of the bacterial immune response, defending against viral infections. In the lab, it has been repurposed as a powerful gene-editing tool. It finds the right gene in the DNA by looking for a matching 20-letter genetic sequence. DNA can be billions of characters in length.  A critical challenge in CRISPR experiment design is ensuring that the system targets only the intended gene, avoiding unintended (off-target) effects.  Our previous research has tackled this by developing specialised data structures and scoring algorithms to estimate the likelihood of off-target activity. Each successive iteration has improved the efficiency of our methods without compromising accuracy, although there remains potential for further enhancements.  The first generation used contiguous k-mers (substrings of length k) to look up, in constant time, other sequences with similar content. For example, AATTCCGG can be broken into its contiguous 2-mers: AA, TT, CC, GG. The position-sensitive 2-mer TT would identify CCTTCCGG, AATTGGAA and GGTTACGG as neighbours, but not TTAACCGG. Specifically, for CRISPR, the 20-character sequence is split into five, position-sensitive 4-mers.  The latest generation shifted the approach from using contiguous k-mers to using longer, gapped k-mer patterns. For example, a gapped 6-mer pattern could be “TT\_C\_AA\_T” (the underscore allows for any character in its place). This gapped k-mer pattern captures a more precise set of similar sequences. Specifically, we used 28 gapped 8-mers. While the size of the k-mer set was larger (from five to 28), it reduced analysis time by 10-fold.  However, the selected set of gapped k-mers is not ideal. We know for certain there is one optimal set, yet it is difficult to find. A naïve approach would find it amongst the 10180 options, but that is an impossible task. So, how do we find it?  This project calls upon those with strong skills in computer science and mathematics to find that solution. You will be hands-on with binary vectors, solving set-coverage problems, and genetic algorithms (those with a search heuristic and optimisation technique). Having background knowledge of relevant biological concepts would be useful but not a necessity. |
| Key words (4-6) | CRISPR, set coverage, data structures, binary vectors, genetic algorithms, optimisation |
| Answerable research questions for 3-5 students (desirable) | This project can be approached from multiple angles, so the research questions remain rather broad but would be narrowed down quite soon after commencing. Here are some to get started:   * How can the optimal set of gapped k-mer patterns be found? * If some accuracy is sacrificed, what is a reasonable set of gapped k-mers that would not risk experimental outcomes? * How can high-performance computing technologies further improve the run-time of the analysis? * How could a hybrid approach be implemented that first approximates risk, and then determines whether a more precise evaluation is necessary? |
| 4-5 key references (desirable) and website resources | 1. Schmitz, C., Bradford, J., Salomone, R., & Perrin, D. (2024, December). Fast and scalable off-target assessment for CRISPR guide RNAs using partial matches. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1649-1654). IEEE. 2. Bradford, J., Chappell, T., & Perrin, D. (2022). Rapid whole-genome identification of high quality CRISPR guide RNAs with the Crackling method. The CRISPR Journal, 5(3), 410-421. 3. Bradford, J., & Perrin, D. (2019). A benchmark of computational CRISPR-Cas9 guide design methods. PLoS Computational Biology, 15(8), e1007274. 4. Bradford, J., & Perrin, D. (2019). Improving CRISPR guide design with consensus approaches. BMC Genomics, 20, 1-11. |
| Required major of studies, desirable skill sets, knowledge, and speciality | * Computer science, mathematics * Strong programming skills in Python and/or C++ * Experience prototyping in a high-level language then reimplementing for high-performance with a lower-level language. * Willingness to learn how to use the QUT HPC, and technologies like OpenMP, MPI, and CPU instruction-sets like AVX512 * Working with Linux systems and version control (git) |
| **Industry-based project: Student IP Agreement.** This is the IP model agreed between the parties. Please note that it is QUT policy that where possible students should be allowed to keep their IP. If students are asked to assign their work, then please **provide a brief rationale** as additional permissions are needed by QUT to approve. | Project IP vests in the student with a license back to Industry Partner **(licence)**  OR  Project IP vests in the Industry Partner/Project owner with a licence back to the student **(assignment)**  OR  Academic project (No IP agreement needed) |
| Number of students (4-5) | Up to 4 |
| The message from supervisor(s) about the acceptance for this project |  |
| Student name(s)  (Print your name and submit this form by the end of Week 2) |  |
| Date |  |
| Remarks on conditions of offer | This is project is one within a broader research effort to improve the CRISPR technology. Participating students will be required to sign an Intellectual Property (IP) assignment with the QUT project owners. |