For the past year, I have been doing bioinformatics research under Dr. John Karro. One significant problem in bioinformatics is the detection of repetitive DNA, which makes up a significant portion of most eukaryotic genomes. My research seeks to improve the sensitivity of an already existing repeat-finding tool, known as Rapid Ab Initio Detection of Elementary Repeats (RAIDER).

In order to quantitatively evaluate RAIDER's performance, I created an evaluation tool that would compare the results of RAIDER and RepeatScout, a widely used repeat-finding tool, when both were run on the same simulated DNA sequence. I created a module to calculate relevant statistics about each tool's performance based on the results obtained. Since the chromosome being used is simulated, the locations of the repeats in the genome are already known. Therefore, it is possible to calculate the number of bases that were accurately and inaccurately categorized as part of a repeat or not part of a repeat. Such information is necessary in order to calculate the sensitivity and specificity for each tool, in addition to other related calculations.

Repeats can be categorized to be either exact or approximate. Over long periods of time, originally identical repetitive sequences have accumulated mutations, creating approximate repeats that are still extremely similar but no longer identical. The detection of approximate repeats can provide information about a particular genome as well as the composition of genomes from previous generations, making a repeat-finding tool's ability to detect approximate repeats extremely valuable. RAIDER, like some other repeat-finding tools, uses spaced seeds to improve the identification of approximate repeats. Spaced seeds are basically patterns describing what positions in two strings must match and what positions in two strings can be a match or a mismatch (don't care positions).

Through my research, I hope to improve upon RAIDER's current ability to detect approximate repeats. I am currently looking into ways to modify the RAIDER algorithm in order for spaced seeds to be used more effectively. I hypothesize that simply changing the primary data structure used in the algorithm could have significant effects on RAIDER's ability to correctly classify approximate repeats.

Additionally, I plan on investigating how to most efficiently use spaced seeds to improve RAIDER's sensitivity. RAIDER currently only allows for a single spaced seed to be used in repeat detection, and this seed is chosen arbitrarily. I will look into the potential use of multiple spaced seeds, as well as investigate which characteristics of a spaced seed make it better suited for repeat-finding purposes. An optimized spaced seed design could allow for a few optimized spaced seeds to be as sensitive to repeat finding as many non-optimized spaced seeds. This information could allow for an improved detection of repeats without significantly affecting the space complexity of the RAIDER algorithm.