# 02-601: Project Proposal: SELEX for DNAzymes

Siddharth Reed, slreed

September 20, 2020

**Problem** The problem is that given some user-defined input DNA $s$ sequence I want to output a potential pool of DNAzyme $D = \{d_1, \ldots, d_n\}$ (catalytic DNA molecules) that will target $s$.

**Scientific/Computational Interest** This is interesting scientifically as DNA is relatively easy to synthesize, both as synthetic oligonucleotides and within cells. Thus being able to design DNAzymes with high affinity to an arbitrary sequence could be useful to easily and specifically cleaving DNA without the need for protein intermediaries, such as CRISPR-Cas9. They may also be useful for researcher who want to study the catalytic properties of DNA and why they have never been found *in vivo*.

**Approach and Feasibility** The approach would be to essentially simulate SELEX *in silico* using a genetic algorithm. One would start with a random pool of sequences and then score them based on whether they would be a suitable DNAzyme for our target (fitness score). Then the best preforming sequences and a new set of sequences "bred" from the highest fitness sequences would form our new sequence pool. This would continue for a set amount of iterations or until the average fitness of the sequence pool remains the same for a set amount of iterations. Since this is a genetic algorithm the breeding would include both mutation and crossover steps for each sequence. The suitability would be based on

- sequence complementarity to the target
- melting temperature of the sequence
- presence of hairpins (palindromes)
- likelihood that a sequence is catalytic

The last element can be estimated by building a machine learning model based on known DNAzyme sequences from the DNArmoreDB Database and other sources that include non-DNAzyme short sequences (aptamers, random sequences etc.). The model (a classifier) would output a probability that a sequence is DNAzyme.

Alternatively the model could evaluate the affinity of the DNA sequence to bind a DNA binding domain, again trained on known binding motifs of the domain such as in a database ENPD or one listed here or a pre-existing tool if it exists.

**External Resources** The sources of the training data are the only external ressources needed, besides computational ressources for training the model and running a genetic algorithm.