Predicting an Aptamer's Target Binding Affinity

Andrew Naber
Electrical Engineering, Stanford University

CS229
Spring 2016

Motivation

Aptamers are short nucleotide sequences that can very specifically bind target molecules. In the process of finding an aptamer that binds a target tightly, several thousand other aptamers are generated and evaluated. This project had two goals aimed at using all of this extra information: **1.** Develop a classifier that would predict whether or not a proposed aptamer would perform well. **2.** Develop a model that can be used to propose new high-performing aptamers.

Methods

1. Support Vector Machine (SVM) Classifier

The dataset was divided into training (60%) and test (40%) subsets. The top 40% of aptamers were labeled as positive examples. Various string kernels were used including spectrum and several variations of mismatch kernels. The k-spectrum kernel measures similarity based on the number of common k-subsequences. Mismatch kernels measure similarity based on spatial nucleotide sequence differences. The optimization was carried out using stochastic gradient descent. The logarithm of fluorescence was used.

2. Gaussian Process Optimization (GP-UCB)

The dataset was treated as the entire domain. That is, following the updates made with a new noisy measurement, the aptamer in the dataset which maximized the upper confidence bound was chosen. The weighted mismatch kernel was used and the logarithm of fluorescence was used.

Data

The dataset was provided by Dr. Tom Soh. It consists of 3,000 different aptamers (nucleotide sequence) and their associated fluorescences (brighter = tighter).

Nucleotide Sequence	Fluorescence
CTCCCGATATCAGAACCTTTTTTGTCACGGAAGGTTGGTCATGGA	1519.0
TGTGTGCTCGCGGCTCGCTAAAGACGCGGTGAAAGGTTGGGTGGT	4039.3
GCGAGTGTGAGGGGATTGGTAGGTGGTTGGCGACCACGGGGGTAG	2320.7

Results: SVM

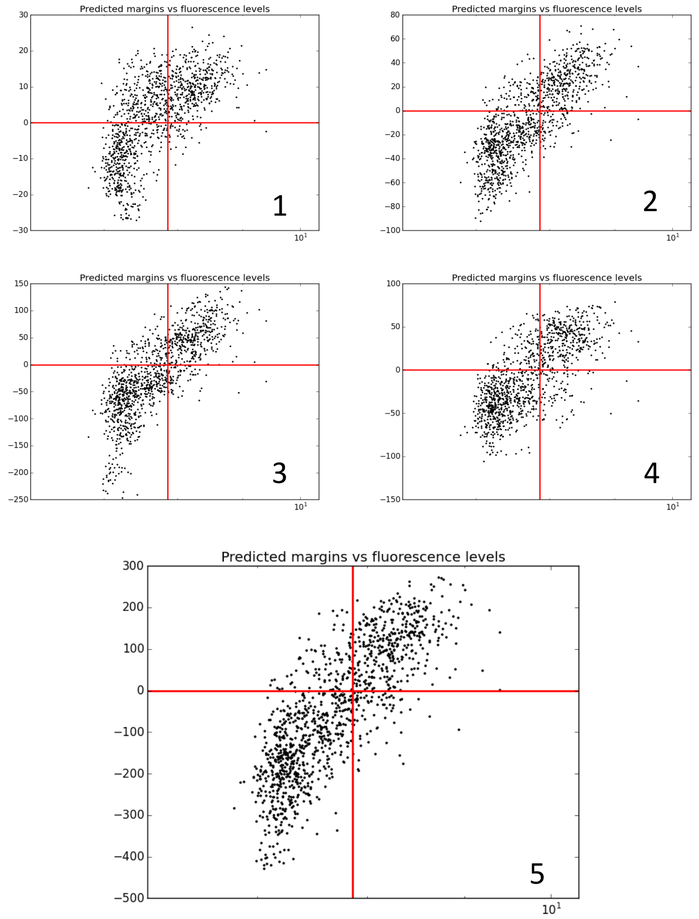
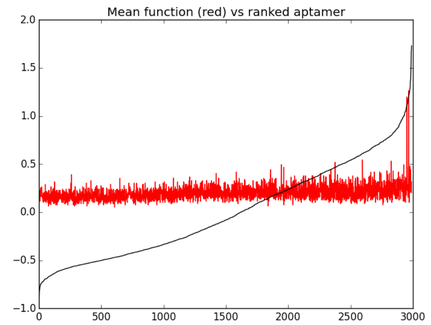
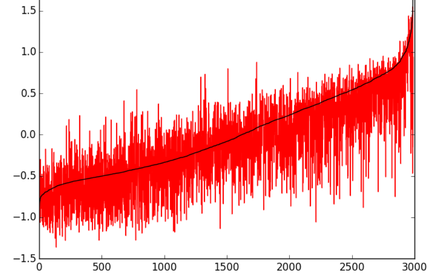


Figure	Kernel	Training Accuracy	Test Accuracy
1	3-Spectrum	76.5%	71.9%
2	3-Mismatch (Chunk)	89.0%	84.5%
3	3-Mismatch (Sliding)	89.3%	85.5%
4	Modified 3-Mismatch (Sliding)	81.3%	83.8%
5	Weighted Mismatch	**89.4%**	**85.6%**

Results: GP-UCB



After First Sample



After 200 Samples

Analysis and Future Work

The SVM classifier performs well with the mismatch kernels, which incorporate spatial information. This makes sense because the location of a nucleotide in the string determines secondary and tertiary folding; some locations will be more sensitive to swapped nucleotides than others. This suggests creating a kernel that weights each location within the string differently depending on sensitivity to mismatches. The GP optimization using UCB results show convergence to the mean function. Future work will combine this with Monte Carlo search methods to propose new high-performing aptamers.