SynBio ML

u6015325

May 6, 2020

Introduction goes here.

1 The need for RBS calculator

1.1 Ribosome Binding Site design

2 Machine Learning Algorithm Description

We optimise the translation initiation rate (TIR) by designing sequential experiments. The goal is to identify the set of RBS sequences with top TIR scores with as fewer rounds as possible.

2.1 Settings

For the first round experiment, we design 180 RBS sequences based on the consensus sequence:

1. 60 RBS sequences designed by "1 by 1 changing" based on the consensus sequence.
2. 60 RBS sequences by random design, including uniformly random (equal probability of choosing each letter for each base); random based on the position probability matrix (PPM).
3. 60 RBS sequences by sequentially machine learning design. We use an bandit optimazation algorithm called Gaussian Process Upper Confidence Bound (GPUCB) [**srinivas2012information**].

2.2 Data Pre-processing

We design the first round bandit recommendation based on the data from **jervis2018machine**. The data contains 113 non-repeated records for 56 unique RBS sequences with the TIR label. The label is between 0 - 100,000 and skewed, which is shown in Figure xx. We normalise the label to 0 - 1 using the min-max normalisation.

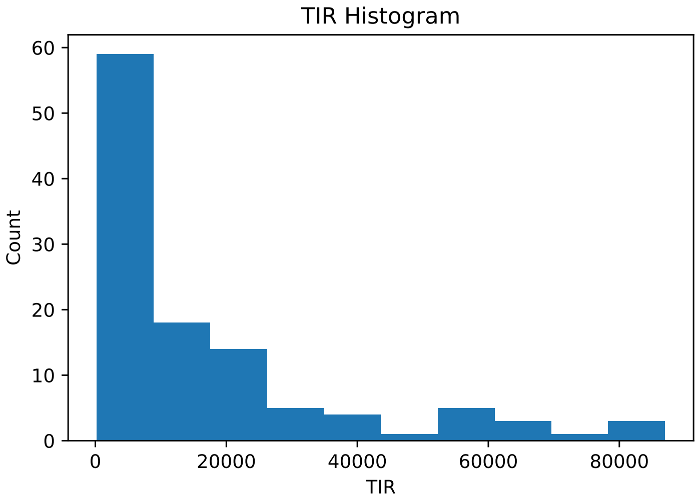

Figure 1: TIR Histogram.

The RBS sequence is 20 bps, we focus on -8 to -13 bps and fix others as the same as the consensus sequence, i.e. TTTAAGA + NNNNNN + TATACAT. For each base, there are 4 possibilities: A, C, G, T. So totally the feature space is $4^6 = 4096$.

2.3 Algorithms

2.3.1 Random sampling

2.3.2 Gaussian Process Upper Confidence Bound

We use Gaussian Process Upper Confidence Bound (GPUCB)] algrithm [**srinivas2012information**], with sum of dot product kernel and spectrum kernel of sequences. The algorithm basically includes two parts:

1. the Gaussian Process regression model, which predicts the label (TIR) and how uncertain we are about our prediction (confidence width); Since the sequences in provided data have the pattern that the core area is different from each other, and other areas are similar. So the kernel for Gaussian Process we are using is the sum of kernels, for core areas we use spectrum kernel with string as input directly, and for other areas we use one-hot encoding and dot product kernel for simplicity.

2. bandit algorithms (Upper Confidence Bound), which recommends sequences to test for next round.

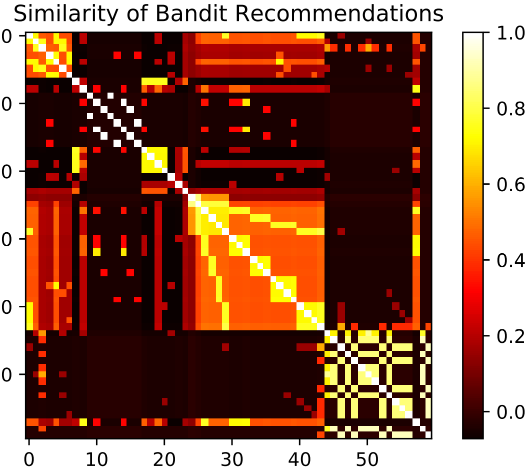

Figure 2: TIR Histogram.

3 Results

1. RMSE of predictions.
2. Similarity of recommendations.