Machine Learning based TIR predictor

u6015325

May 7, 2020

1 Introduction

One of the main tenets of synthetic biology is design, evaluation and standarization of genetic parts [**Brophy2014**, **Canton2008**, **Stanton2014**]. This can be done in variety of ways, most of each involve designing the DNA sequence in CAD software and then physically testing it in a laboratory. Alternative to this is computer modelling and prediction of part behoviour based on the designed DNA sequence or design of DNA sequence based on expected function [**Yeoh2019**, **Nielsen2016**]. Most of these models are based on either the thermodynamic properties of the involved molecules (DNA, RNA, proteins among others) or empirically obtained values describing a relevant to design value, like Translation Initiation Rate (TRI) in case of Ribosome Binding Sites (RBS) [**Xia1998**, **Chen1994**, **Vellanoweth1992**, **Chen2013**, **Reeve2014**].
According to Reeve *et al.* there are three main RBS calculators, all prediciting the TRI based on the thermodynamic properties of the RBS and the ribosome [**Seo2013**, **Na2010**, **Salis2009**]. Predictions from all of these models are relatively good ($R^2 > 0.8$), there come with a number of caveats: i) they rely on calculations of free energies that can be hard to calculate ii) in general, the models' accuracy is improved by increasing the number of phenomenons taking place during the translation, but this can lead to paradoxically decreased model accuracy due to accumulation of errors [**EspahBorujeni2016**] and iii) by using deterministic coefficients to calculate energies one disregards often stochastic nature of processes in the cells which again increases perceived prediction error [**Goss1998**].
Synthetic biology is currently going through a phase of exponential increase in volume of data produced during experiments. [**Freemont2019**] New experimental methods heavily relying on advances in automation and microfludics allow unprecedented precision and throughputs in data generation. This
In this article we present the first in our knowledge TIR predictor built using machine learning approach.

2 Ribosome Binding Site design

3 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.

3.1 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**].

3.2 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.

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 $4^6 = 4096$.

3.3 Algorithms

3.3.1 Random sampling

3.3.2 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

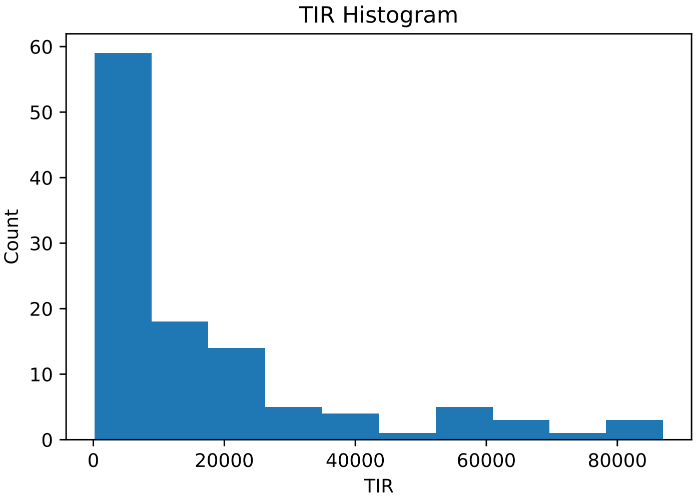

Figure 1: TIR Histogram.

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.

4 Results

1. RMSE of predictions.

2. Similarity of recommendations.

5 Methods

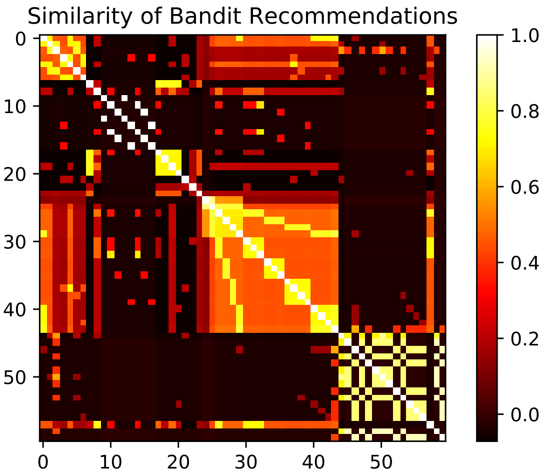

Figure 2: TIR Histogram.