rRNA-mRNA interaction subsequence optimization model

Dept. Bio-Medical Engineering, Univ. Tel-Aviv, Israel.
Boaz Apt, Shir Bahiri Elitzur, Tamir Tuller (tamirtul@gmail.com)
Project Git: ***If we will add the took then we will upload to git***

Abstract

The problem we investigated belongs to the world of bioinformatica. The goal was to examine whether there are creatures that uses different sequences than the well known Shine Delgarno sequence. At a high level, we want to find such sequences, based on the creature genes full sequences, that will result in higher PA levels then the canonical SD.

1. Introduction

What we wanted to investigate is whether we can build a model that will find a sequence that gives us the lowest hybridization energy and validate this against a common prediction index that researches uses called CAI

We believe that a tool that gets a creature genum and provide us with a 6 nucleotide sequence, if exist, that result in high PA, can help managing disease, help in genetic engineering ect.

For data we search for bacterias that have fully and complete sequence and process them to extract the data we need.

Obviously it was better to get true protein abundance but this data is hard to achieve and not in the scope of this initial work. To validate this decision, we checked the correlation of the CAI index with true PA levels of e. Coli and got correlation of 0.43 with p-value of 3.35e-184 (Fig 1.1)

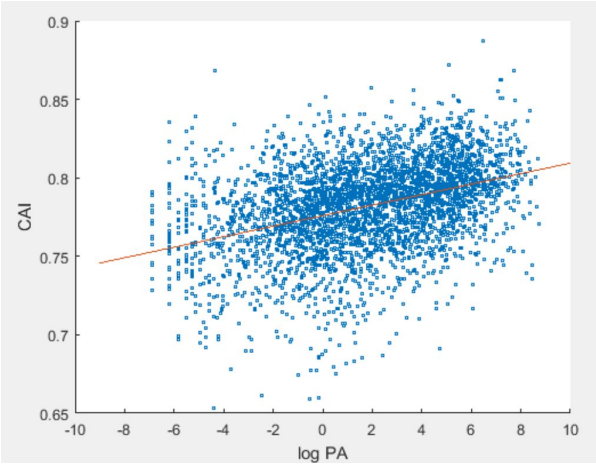
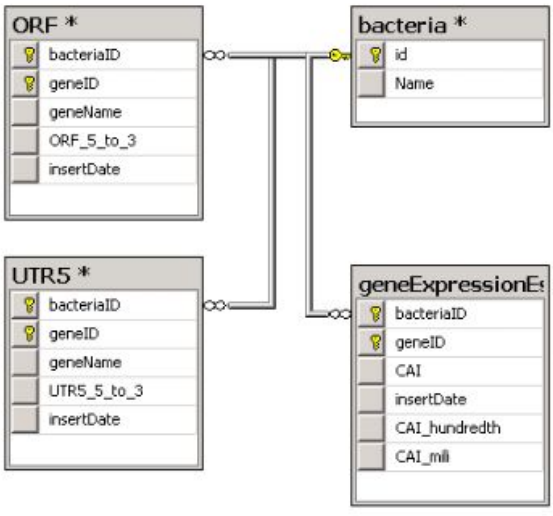


Fig 1.1

2. Dataset

The dataset is generated based on 600 bacteria from the following phyla or classes: Alphaprobacteria, Betaprotobacteria, Cyanobacteria, Delataprotobacteria, Gammaprtobacteria, Gram positive bacteria, Purple bacteria, Spirochaetes bacteria. All of the bacterial genomes were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/). For bacteria we extract all the genes **O**pen **R**eading **F**rame (the coding region) and we also extract 50nt upstream of the start codon (Which is the end of the **U**n**T**ranslated **R**egion).

The data was uploaded into relational DB (Fig 2.1)



During the data processing we have clean some data and finally we ended up with 551 bacteria, each bacteria have around 4000 gene so we have tataly 2,233,121 row data

next

Another dataset

3. Method

iterative algorithm, in each iteration we tries to find an aSD sequence (“suggested aSD”) that both significantly improves the average hybridization energy between the rRNA-UTRs and also improves the correlation with DCBS (quantitative method

of predicting the level of expression)

Step 1: Extract possible aSD sequences from rRNA:

```
AGGCGACGTGCCGTTCTAGGCCGGCCCGCGAC
AGGCGACGTGCCGTTCTAGGCCGGCCCGCGAC
AGGCGACGTGCCGTTCTAGGCCGGCCCGCGAC
AGGCGACGTGCCGTTCTAGGCCGGCCCGCGAC
```

Possible aSD: CGTGCC, GTGCCG,

TGCCGT, GCCGTT…….

Step 2: For each possible aSD (“current possible aSD”), we calculate the average hybridization energy :
2.1 loop all the bacteria gene’s UTR5:

	1	2	3	4	5	6	.	.	.	50
UTR_1	A	C	C	T	G	G	.	.	.	A
:	:	:	:	:	:	:	:	:	:	:
UTR_n	C	G	T	A	A	T	.	.	.	T

2.2 extract “possible SDs” for UTR and Loop on them
2.3 Loop on both “current possible aSD” and “already suggested” aSD (=”suggested aSD” from previous Iteration, if exist)
2.4 Calculate hybridization energy

This is what we get in this stage:

	Gene 1	Gene 2	Gene 3	...	Gene N	Average H.E.
aSD GCCGTT(1) & CGTGCC(2)	-1.9	-2.8	0.9	...	-6.1	-2.1
aSD GCCGTT & GTGCCG	-0.1	0	-8	...	-3.1	-3.7
aSD GCCGTT & TGCCGT	-0.4	-2	-8.4	...	-9.2	-4.37

(1) is already suggested aSD sequence/s from previous iteration
(2) is possible aSD to add in this iteration

Step 3: Select the aSD sequence that brings the average hybridization energy to minimum. In the above example - TGCCGT

Step 4: Compare the average hybridization energy to the previous iteration. If significantly improves then we continue, otherwise we stop

Step 5: Calculate energy vector with permut UTR5 and run wilcoxon test

Step 6: Check that the correlation of the energy vector with CAI and DCBS improved

Step 7: Add the aSD of this iteration to the

“suggested aSD” and start over from step 1.

מסמך שאני עובד איתו:
https://docs.google.com/document/d/1jChvPBaWde7D1_5Sfu58lJB3V96NZR67gFwfVlLtMq0/edit