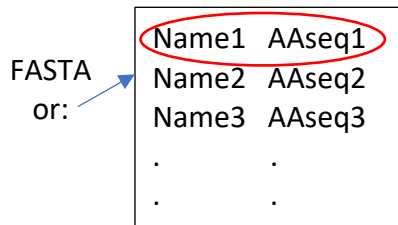


Algorithm Flow

Input File



Save entries in
dictionary
Name:AAseq
(Key:Value)

Loop over
dictionary
Seq by Seq.

Start
loop

*
'AAseq1' 4

Loop over
AAseq

Get codons
and weights

(from single codon usage
table or CC table)¹

Codons = ["TTT", "TTC"]
Weights = [33, 67]

GC correction

-if len(newSeq1) > 10: measure
GC content.
-Correct weights (4P logistic)

NewWeights = [50, 50]

-Screen for last synonymous
codon within the previous 25
codons.
-correct weights accordingly

NewWeights = [60, 40]

Autocorrelation
Bias correction

Randomly select
codon. Add to
the newSeq

newSeq1 = 'TTT'

Check for
restriction sites
and other
motifs.
Cut if necessary.

newSeq1 finished

Until seq
back-translated

x10

GC content within range
(48-60%)

GC content NOT within
range (48-60%)

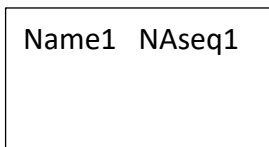
-Calculate candidate score².
-Store in temporary
dictionary.
-Go back to * x10 (for ten
candidates)³.

Select and save the candidate
with highest score in output
file.

-Delete Seq. Back-translate again. newSeq1 = " *

-Every 10 times:
-Relax the respective
threshold by 0.5%
MaxThreshold = 60.5%

Output File



Go to the next
AAseq

*
'AAseq2'

All the
sequences
completed

End of loop

Print output if
desired

End of script

Algorithm Flow

NOTES:

1) CC applies only for the bicodons with significantly different counts in highly expressed genes compared to low- and the full transcriptome. The single codon usage table is used for the rest (About 85% of the times for B-cells. For HEK this is lower).

2) SeqScore = sum(C.A.I, GC_score, CpG_score)

The GC_score and CpG_score are negative numbers. C.A.I = Codon Adaptation Index.

3) The process of backtranslating 10 candidate sequences is done concurrently. Depending on the cores of your machine, 4 or more candidates can be backtranslated at the same time.

4) For every AAseq to be backtranslated: 10 candidate sub-strings of the first 20 AAs are generated. The minimum free energy is calculated and the one with the highest is chosen. This 60-nucleotide long sequence is used as a starting point to make the 10 candidate full-strings.

Rationale: The sequence with the highest minimum free energy should be the one forming the least thermodynamically stable secondary structure. This in turn should favor translation initiation.