Why this one? The influence of nucleotide context on synonymous codon choice

Natasha Mikhaylovskaya, Alexey Kolodyazhnyy, Eve Zubova,
Asya Mendelevich, Eugenia Khodzhaeva, Valya Burskaia, Zoe Chervontseva

Abstract

Is it possible to predict particular synonymous mutation, based on adjacent 100 nucleotides?

We used convolutional neural network to find patterns that determine synonymous evolution (in particular position). We found two entities that determine direction of synonymous substitution: it is the **state of the next codon** and the **states of 3rd codon positions** in the whole alignment. Surprisingly, the model trained on one amino acid, works great for some other amino acids.

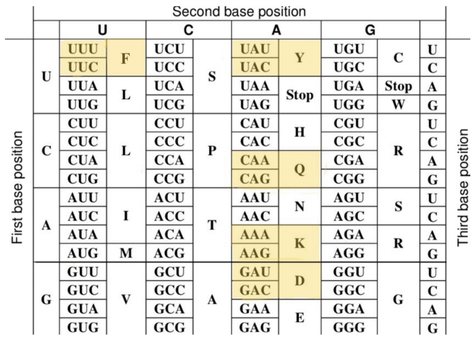

Fig.0: Genetic code. The considered amino acids and their codons are marked in yellow.

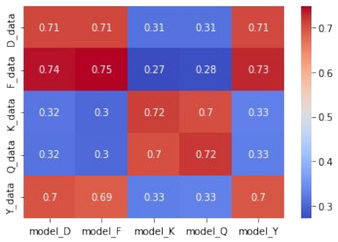

Fig.1: Cross evaluation of various models. Models seem to work equally well for various amino acids even if not trained on them. The figure shows that models for amino acids D, F, and Y get good results on each other's datasets, as well as models for K and Q. This is a consequence of the fact that in the first three datasets, we analyzed T/C variants and in the last two, we took A/G variants (see the **Genetic Code**). But also, we can see that inversion of predictions of models allows us to get good results on groups with different 3rd position variants. That means that models may have learned similar patterns.

Data

aa	train_size	test_size	baseline_accuracy
E	182132	19914	0.61
K	161506	17788	0.65
F	172562	19032	0.69
D	189290	20784	0.63
Y	146258	15820	0.63
H	120546	13352	0.64
Q	136860	15048	0.69
N	161642	17946	0.65
C	62256	6970	0.66

Table 0: Train and test datasets. The datasets consist of the non-overlapping non-homologous nucleotide sequences of +-48 nt around the codon of interest. All the sequences were cut from the protein coding genes of bacteria. Every dataset was balanced to have equal number of contexts for each codon. Baseline accuracy was estimated by NaiveBayes.

Methods

Neural network architecture:

- Convolutional layers, Multi-Head Attention, Bidirectional LSTM
- Relu activation
- Dropoff regularisation

Results

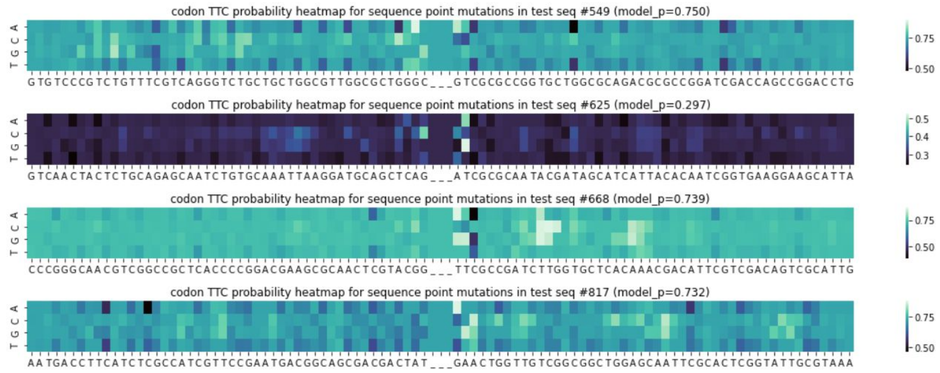

Fig.2: Example plots show the dependence of the predicted codon probability (unit color) on the point mutations (rows) in the original sequence. *For this figure we used 4 random contexts of Phenylalanine (F).*

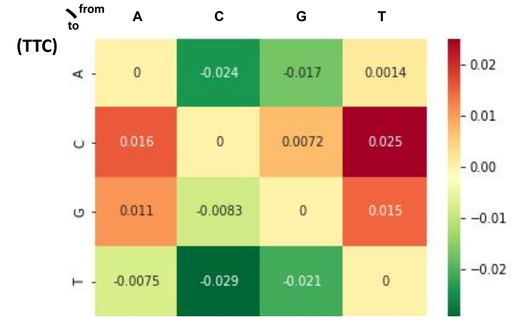

Fig.3: Average changes in probabilities of prediction of the TTC codon (F) after replacing the original nucleotide (y-axis) with a mutant one (x-axis). The probability of correct codon prediction is higher when A or T mutated to C or G, than vice versa. For TTT, on the contrary, mutations from A or T to C or G reduce predictability power.

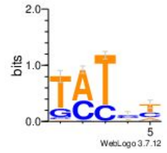

Fig. 4: Logo for the one of important convolutional filters.

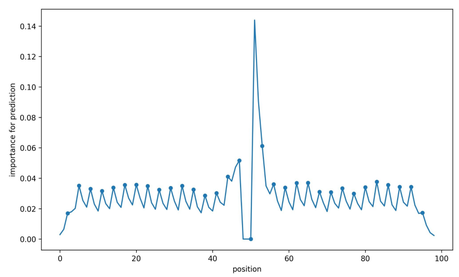

Fig.5: Average effect of mutations for each position in a context. The figure shows that 4 positions in advance (45-48) and especially 4 positions just after (52-55) are the most important for codon predictability, and that 3rd positions in along the entire length of a context are more important than 1st and 2nd.
Large points depict 3rd positions in codons.

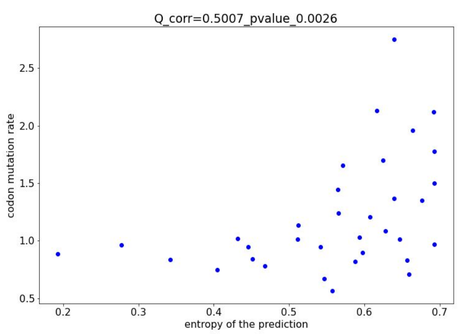

Fig.6: Uncertainty level of the neural network (x-axis) vs. variability of Glutamine encoding codon (y-axis). The more variability the codon has, the more uncertain neural network is. The results show the significant (p-value < 0,05) correlation. However, such correlation is detected only for Glutamine (Q).

Conclusions

- For F, Y, Q, K, D we got models that work better than the Bayesian baseline, which considers all context positions independent. However, the resulting accuracy is only ~5% higher than the baseline hence needs further improvement.
- Our models seem to work equally well for various datasets even if not trained on them.
- The most important entities for choosing synonymous variant are the state of the next codon and the states of 3rd codon positions in the whole sequence.
- At first sight, the possible nucleotide pairing in RNA secondary structure with the codon of interest doesn't have significant influence on choosing synonymous variant (Data not shown).
- Average prediction confidence for TTC over TTT rises if adenines and thymines around the site of interest are replaced by guanines or cytosines.