Computer Analyses of Nucleotide Patterns around Start Codons

Rintaro Saito
rsaito@mag.keio.ac.jp

Hidekazu Sasaki
s94197hs@sfc.keio.ac.jp

Yuko Osada
t94092yo@sfc.keio.ac.jp

Masaru Tomita
mt@sfc.keio.ac.jp

Laboratory for Bioinformatics

Graduate School of Media and Governance

Department of Policy Management

Department of Environmental Information
Keio University
5322 Endo, Fujisawa, 252 Japan

Translation is a process that ribosomes synthesize protein by scanning a mRNA from left to right. Although ribosome initiates translation at the first found AUG trinucleotide in eucaryotes most of the time, it sometimes skips and ignores the first AUG and initiates translation at the second or third AUG trinucleotide. The exact mechanism of selecting AUG for translation initiation is not known.

We first conducted comprehensive computer analyses of nucleotide, dinucleotide and trinucleotide distributions around start codons and around skipped AUGs. Previously, Kozak[1] conducted similar analyses with 699 mRNA translation initiation sites. Our analysis data were taken from the entire Genbank database(Release 94.0), which contains about ten thousand mRNA sequences altogether. Sequences of mitochondria and immunoglobulin are excluded from our analyses because of their special translation initiation features.

Profiles of nucleotide distribution around start codons were constructed and the entropy for each position was calculated.

The results for 4 bacteria are shown in Figure 1. The entropy value is sharply low at the position -11 in *Bacillus subtilis* and position -9 in *E. coli* and *H. influenzae*, respectively. The sequences around the positions are presumably Shine-Dalgarno sequences. The curve representing *Mycoplasma genitalium*, however, does not go down as sharply at these positions.

Secondly, we investigated the cases where two AUG trinucleotides are located close to each other. While the first AUG is generally selected as a start codon for the most of the time, we

found that the second AUG is frequently selected (by skipping the first) if the two AUG's are separated by exactly 2 base pairs.

Thirdly, we analyzed frequencies of stop codons located downstream of the skipped AUGs. It is known that first AUG is skipped if followed immediately by a stop codon[3]. Figure 2 shows that the stop codon frequency gradually decreases until about position +30 in both vertebrates and invertebrates. Also, there are periodic peaks of the frequencies every 3 base positions from the skipped AUG. In other words, stop codons exist preferably in the same reading frame as the skipped AUG.

Finally, we analyzed frequencies of AUG trinucleotides that appear upstream of start codons. The results show that there is a general tendency to have fewer AUG trinucleotides in front of start codons(Figure 3).

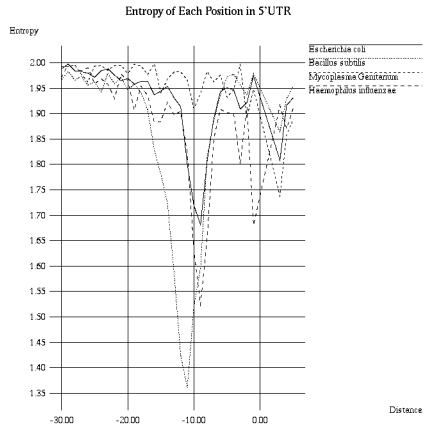

Figure 1

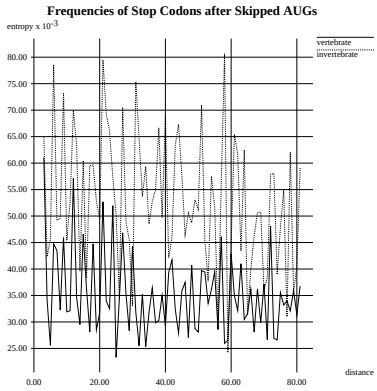

Figure 2

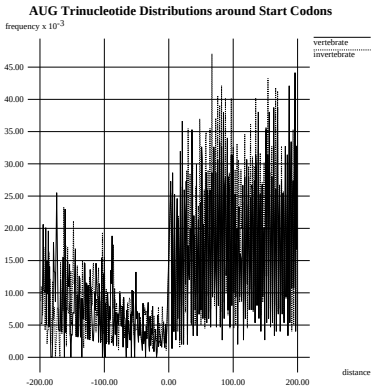

Figure 3

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from the Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Research 15:8125-8148

[2] Shine, J. and Dalgarno, L. (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. Proc.Nat.Acad.Sci.USA 71:1342-1346

[3] Kozak, M. (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. Mol.Cell.Biol. 10:3438-3445