DNA evolved to minimize frameshift mutations

Valentina Agoni  
valentina.agoni@unipv.it

Point mutations can surely be dangerous but what is worst than to lose the reading frame?! Does DNA evolved a strategy to try to limit frameshift mutations?! Here we investigate if DNA sequences effectively evolved a system to minimize frameshift mutations analyzing the transcripts of proteins with high molecular weights.

During replication the DNA polymerase creates an error every $10^5$-$10^6$ nucleotides [1] [2]. In [3] we demonstrate that the amino acids coding table evolved to minimize point mutations. They can of course be dangerous but what is worst than to lose the reading frame?! So how to limit frameshift mutations? To verify if the nucleotide sequences (not the amino acids coding table as for point mutations [3]) evolved to minimize the probability of frameshift mutations we analyzed the amino acids coding table to find the proper sequences to consider.

We are interested in a sequence the more homogeneous as possible in terms of nucleotides composition to compare its frequency with a less homogeneous one. First there are of course 4 codons composed by the repetition of the same nucleotide: AAA, CCC, GGG and TTT (blue circles in Figure 1). If we look at the amino acids coding table we notice that 3 of them code for amino acids that are associated also with many other triplets. On the contrary we want to compare the frequency of 2 sequences undergoing differences due to the fact that the associated amino acids could be coded by a diverse number of codons.

| First Position | Second Position | Third Position |
|----------------|-----------------|---------------|
| **U**<br>CCU  | **C**| **G**<br>CCU  | **A**<br>UAA  | **C**<br>GAG  | **U**<br>GGU  |
| UUA  | UCA  | UGC  | UAU  | UCG  | UGU  |
| UUC  | UCC  | UCG  | UAC  | UGG  | UGC  |
| UUA  | UCA  | UGC  | UAU  | UCG  | UGU  |
| UUC  | UCC  | UCG  | UAC  | UGG  | UGC  |

[Figure 1: Amino acids coding table with selected codons marked with blue circles.]
**Figure 1. The amino acids coding table.** We are interested in finding 2 codons constituted by the repetition of the same nucleotide for the higher number times. The amino acids coding table reveals that for our purpose we cannot use Phe and Leu because it can be coded by many codons. On the other hand the Lys-Arg duo is perfect.

We cannot compare for example the frequency of AAA with the frequency of GGG because they are affected for example by the amino acids frequency and by the number of codons associated with a single amino acid. Therefore we focalized on Lys (K) that can be codified by AAA or AAG. Moreover a single codon is not representative to check the frameshift probability because we know that a shift of DNA polymerase is much higher the longer is the sequence composed by the repetition of a single nucleotide. The probability of a pre-extablished amino acids triplet was too low even in long transcripts.

In particular we select the amino acids pair LysArg (KN) because:

- there is only one other amino acids codified only by a similar sequence of nucleotides: Asn (N) that can be codified by AAU or AAC
- both these 2 amino acids could be coded only by the 2 codons under investigation
- moreover to study an amino acids couple allows us to eliminate problems related to their different frequency in the DNA.

In this way the comparison will be balanced. Hence the best thing to do is to consider the couple KN namely to check the frequency of AAGAAC and AAGAAT respect to the frequency of AAAAAAC and AAAAAAT.

We analyzed proteins with a high molecular weight (that means with long transcripts) in order to have a great number of KN. This number is obviously influenced by Single Nucleotide Polymorphisms (SNPs) and by alternative splicing. It is useful to clarify that we analyzed the frequency of 2 amino acids and into mRNAs despite the principle of frameshift mutations minimization is relative to DNA replication for the reasons listed before. If the nucleotide sequences evolved to minimize the probability of frameshift mutations we should find AAGAAC and AAGAAT (green in Figure 2) more than AAAAAAC and AAAAAAT (yellow). This is what we actually found. Figure 2 reports the sequences analyzed: human myosin [4], human BRCA1 (breast cancer type 1)[5] and murine 53BP1 [6]. We evaluated the ratio between green sequences and green plus yellow sequences. The results are respectively: 1, 0.7 and 0.6 with an average of 0.77. Comparing it with the equal probability to have green and yellow sequences (ratio: 0.5 with a negligible standard deviation) using a one tailed t-Student’s test we have a p value of 0.04537.
Figure 2. mRNA sequences of human myosin, human BRCA1 (breast cancer type 1) and murine 53BP1 from NCBI [4] [5] [6]. If the nucleotide sequences evolved to minimize the probability of frameshift mutations we should find AAGAAC and AAGAAT (green) more than AAAAAAC and AAAAAAT (yellow). The vertical lines reflect the partition into amino acids to highlight that we do not consider for example all AAAAAAC sequences but only the ones corresponding to KN, excluding for example --A AAA AC- which is inevitable despite shifts.

Bibliography

1. Griffiths, William M.; Miller, Jeffrey H.; Suzuki, David T.; Lewontin, Richard C.; Gelbart, eds. (2000). "Spontaneous mutations". An Introduction to Genetic Analysis (7th ed.). New York: W. H. Freeman. ISBN 0-7167-3520-2.

2. Freisinger, E; Grollman, AP; Miller, H; Kisker, C (2004). "Lesion (in)tolerance reveals insights into DNA replication fidelity". The EMBO Journal 23 (7): 1494-505. doi:10.1038/sj.emboj.7600158. PMC 391067. PMID 15057282.

3. Agoni, V. (2013) “Insights Into Quantitative Biology: analysis of cellular adaptation”. arXiv:1307.0451

4. http://www.ncbi.nlm.nih.gov/nuccore/558668?report=genbank

5. http://www.ncbi.nlm.nih.gov/nuccore/237757283?report=genbank

6. http://www.ncbi.nlm.nih.gov/nuccore/16754835?report=genbank