

Poster A-10

Codon instability at high GC level in the comparison between Human and Chimpanzee



Authors:

Rosina Piovani (*Laboratorio de Organización y Evolución del Genoma. Sección Biomatemática. Facultad de Ciencias, Uni*)

Hugo Naya (*Laboratorio de Organización y Evolución del Genoma. Sección Biomatemática. Facultad de Ciencias, Uni*)

Victor Sabbía (*Laboratorio de Organización y Evolución del Genoma. Sección Biomatemática. Facultad de Ciencias, Uni*)

Héctor Musto (*Laboratorio de Organización y Evolución del Genoma. Sección Biomatemática. Facultad de Ciencias, Uni*)

Short Abstract: We obtained a set of 2482 orthologous genes between Human and Chimpanzee and analyzed the codon usage of sites where the amino acid was conserved. We found that as long as GC3 increases the level of conservation decreases for most amino acids. We discuss the results as a consequence of natural selection.

Long Abstract:

Over the past ten years there has been a huge increase in the amount of completely sequenced genomes that made molecular evolutionists to start thinking in a higher level, the genomic level. That made possible studies like comparing rates of genome evolution, finding orthologous genes among genomes, timing genomic divergence and so on. All these studies are progressively making it possible to understand and analyze the evolutionary patterns among species, and also the compositionally trends that they might have undergone. This has been a very important step because, as said before, it is a way to study this features in another level, not the genes nor the proteins but the whole genome. Taking this into account and focusing on the human genome, this study aims to analyze the trends in codon usage in this genome in comparison with its most related one: the chimpanzee genome. The first step to approach this study was to identify orthologous genes between the two genomes. The performance of reciprocal blasts led to a set of 2482 orthologous genes which were used for the whole study. In order to start the analysis on codon usage between both species we prompted the discussion concerning when the amino acid was conserved in the set of orthologous genes. If that was accomplished we wanted to know weather the codon was also conserved. There are several papers that claim that the differences between codon usage strategies in vertebrates are due to differences in genome organization, due to the presence of isochores. These are fragments of the genome that have a relatively constant GC content, so according to these studies the differences in codon usage might be only due to the localization of the genes in these regions. To avoid, what we could call, this "isochore effect", and since we are studying synonymous changes (changes in codons that do not change the amino acid); we separated all the orthologous genes into several groups taken every 5% of GC3 content and analyzed the eleven resulting windows separately. By doing this, the GC content of the third codon position (GC3), which is the greatest influence in synonymous changes, has less influence in the codon choice, making this situation ideal to identify other features influencing codon usage. We wrote a script to calculate all the synonymous and non

synonymous changes per codon. Using this script we calculated what we called the “coincidence coefficient” referring to the ratio between the conserved codons and the total codons for each amino acid. This is a very useful way to analyze this type of changes because it is different from estimating a synonymous substitution distance in the sense that we are not comparing at the gene level but at the amino acid level, which allows us to better understand the codon usage trends. When we plotted this coincidence coefficient against the average GC3 of all the windows we observed that all the amino acids had a negative correlation which in fact was significant in thirteen out of eighteen amino acids. That means that synonymous changes are more frequent in GC3-rich genes than in GC3-poor ones. The amino acids with the most significant correlation were Ser, Pro, Ala and Thr. Surprisingly, they are all the amino acids that have C in the second codon position and they are also part of quartets, so they can form the CpG dinucleotide in the second and third codon position. We analyzed the frequency of the NCG codons in these amino acids with respect to the others and we compared it with other amino acids with A, T or G in the second codon position. We found that in the GC3-rich genes the NCG codon is used less than expected, and the results suggest that this is due to the known effect of CpG avoidance, since this dinucleotide is very susceptible to deamination of 5-mC to T which leads to CpA and TpG if it is not repaired. This could explain, for these four amino acids, the presence of more synonymous changes in the GC3-rich genes, at least in the transition NCG to NCC. Furthermore, we analyzed the conservation of the amino acid subtracting the number of conserved residues by the total of amino acids without considering the cases in which it was paired with an indel in the alignment. We plotted the frequency of conserved amino acids (calculated for each window) against the GC3 content of genes and we found a negative correlation. This result indicates that the instability observed at higher levels of CG3 at the codon level is also seen at the amino acid level. In order to get a more detailed picture, we calculated the correlation coefficient in each codon instead of amino acid (the fraction of conserved codons over the total number of codons for a given amino acid). We plotted these values (calculated for each GC3 window) against the number of copies of isoacceptor tARN for each codon and we found a positive correlation in each window. This means that the most conserved codons are the ones that are recognized by the isoacceptor tARN with higher copies. This implies a tendency to conserve the codons that guarantee a better speed of translation and/or accuracy. One interesting feature of this result is that the plot correlation is more significant in CG3-rich genes, which suggests that these genes could have a higher level of expression which would explain the conservation of the major codons. To sum up, using a set of orthologous genes between Human and Chimpanzee we found that genes with higher GC3 content have a lower level of amino acid conservation but when it is conserved the number of synonymous substitutions of codons is higher. The conserved codons are probably fixed by the action of natural selection acting at the level of translation (accuracy and/or speed).