Exploring the evolution of coding sequences in metazoan mtDNAs: effects of mutation and selection in Insects and Vertebrates

Castellana S[1], Vicario S[2], Donvito G[3], Saccone C[4]

Motivation

In prokaryotic and eukaryotic nuclear genomes, selection on synonymous variability of coding sequences has been shown; synonymous codons that allow an accurate and rapid translation of transcripts, are used preferentially. For example, the cellular intolerance to misfolded proteins, induced by incorporation of uncorrected amino acids (a.k.a. MIM or 'Mistranslated-Induced Misfolding'[1]), may be a major cause of codon usage bias. We investigated the synonymous variability on a large set of mitochondrial genomes, trying to put in evidence the differential role of mutational and non-mutational factors that shape variability in codon usage. This could be a useful contribution to the ongoing debate [2,3,4,5] on the prevalence of adaptation in the evolution of mitochondrial genome. Then, we compared synonymous and non-synonymous variability to determine a possible correlation between them, as can be seen in the context of selection by MIM. Our data provide new elements of discussion about the evolutionary history (and its implications) of metazoan mtDNA.

Methods

We collected over 1,000 and over 100 RefSeq mitochondrial genomes from Genbank, belonging to sub-phylum 'Vertebrata' and class 'Insecta'. Positional and base compositional information were calculated for all coding sequences. We use two approaches to analyze the codon usage in these genes. At first, we perform a linear regression analysis to determine the contribution of different predictors to our variable of interest, the index ENC ('effective number of codons'). This statistics gives a measure of the bias in synonymous codon usage: it ranges from 20 to 61 (maximum bias, i.e., one codon for one amino acid to minimum bias, i.e., all codons are used). Predictors include 'Species', 'Gene', 'Position', 'Strand', 'expected ENC': this last variable has been estimated by base composition of the third positions of quartets (Val, Pro, Thr, Arg, Ala, Gly codon families) and it can be considered as a good estimation of codon usage bias, in case of the exclusive contribution of mutational forces. Then, to overcome the shortcoming of an index-based approach (impossibility to appreciate differences in variance due to low

[1] Dipartimento di Genetica e Microbiologia, Università degli Studi di Bari, Bari. [2] Consiglio Nazionale delle Ricerche - Istituto di Tecnologie Biomediche, Bari. [3] Istituto Nazionale di Fisica Nucleare, Bari [4] Dipartimento di Biochimica e Biologia Molecolare "E.Quagliariello", Università degli Studi di Bari, Bari

counts among different genes and codon families) we tested few key results of previous studies in linear model analysis using a likelihood approach on codon usage and base composition for the cited codon families. The observed differential synonymous codon usage has been treated as a multinomial problem in which the relative occurrence probabilities of each synonymous codons have been inferred by different ways of modeling codon counts Each models assume that codon choice is guided by genome-wide forces (i.e. MIM) or gene specific ones under the control of base composition or not. Likelihood model comparisons have been performed by Likelihood Ratio Tests: models were evaluated globally by overall p-value and specifically by Bonferroni sequential correction for multiple testing. In the following step, we calculated the non-synonymous and synonymous substitution rates for gene multialignments relative to about 1000 genera of Vertebrata by using: Tranalign (EMBOSS package) to build nucleotidic multialignments; mrBayes (GTR+G model) to obtain phylogenetic trees; Codeml (PAML package) to calculate non-synonymous and synonymous substitution rates for each multialignment.

Results

We conclude that effective number of codons is significantly dependent on 'expected ENC', 'Species' and 'Genes'; on the other hand, the expected effective number of codons well resumes the other factors such as gene location (related to replication origin) and strand. The 14 and 15% of ENC variability (for Insecta and Vertebrata, respectively) is associated directly to gene and species and not to gene specific base composition. About the mean effect of gene on ENC, there is a common trend in 'Insecta' and 'Vertebrata', but three genes (ND1,ND4,ND5) have a contrasting behavior in the two groups. Base composition likelihood models confirmed the gene-specific mutational input in vertebrate and insect genome: this could be linked to the asymmetrical way of replication of mtDNA. For codon usage models, the model in which codon usage is constant among genes but determined by the optimality of each codon type, is always rejected. Indeed, the model in which codon usage is inferred by mutational gene-specific input is rejected at 'species' level by 20 and 40% of the Insecta and Vertebrata species. This means that, although the mutational forces play a major role in determining the codon variability, in a remarkably large subset of the two group of data they cannot be sufficient. Non-synonymous and synonymous substitution rates present a very dis-homogenous pattern across the different genomes, which shows a differential impact of mutation and selection for optimal codons in the different taxa.

Contact e-mail

stefano199@gmail.com, saverio.vicario@ba.itb.cnr.it

Supplementary information

[1] Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell, 134, 341-

352. [2] Galtier, N., Bazin, E., & Glemin, S. (2006). Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals. Science, 312(April), 570-572. [3] Nabholz, B., Bazin, E., Galtier, N., & Glemin, S. (2008). Determination of Mitochondrial Genetic Diversity in Mammals. Genetics, 361(January), 351-361. [4] Meiklejohn, C. D., Montooth, K.L., & Rand, D.M. (2007). Positive and negative selection on the mitochondrial genome. Trends Genet., 23, 259-263. [5] da Fonseca, R. R., Johnson, W. E., O'Brien, S. J., Ramos, M. J., & Antunes, A. (2008). The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics, 9(119).