Evolution of Codon Usage Bias in *E.coli*

Fanny Pouyet[1], Marc Bailly-Bechet[1], and Laurent Guéguen[1]

[1]Laboratoire de Biologie et Biométrie Evolutive
University Claude Bernard Lyon 1
43 bd du 11 novembre 1918 69622 VILLEURBANNE cedex
{fanny.pouyet, marc.bailly-bechet, laurent.gueguen}@univ-lyon1.fr

December 12, 2013

Abstract

We develop an evolutionary model, inspired from Yang and Nielsen [1], that takes into account evolutionary processes at nucleotidic, codons and amino acids levels. It is implemented in Bio++ suite [2]. We apply this model in an homogeneous, non-stationary context. We study evolution of codon usage bias, preference of synonymous codons over others, in three close strains of *E. coli*. It computes equilibrium and root frequencies. We show that nucleotidic mutational bias and codon bias are counteracting: mutational bias tends to increase AT composition whereas codon bias favors GC enrichment.
Keywords : Evolutionary Model, Codon Usage Bias, *Escherichia coli*

Introduction

The genetic code is redundant, and some codons, called synonymous, are translated in the same amino acid. Degeneracy of the code does not however lead to a uniform usage of those synonymous codons: at species or at genomic level, a particular codon is preferrentially used over other synonymous ones [3]. Codon usage bias may vary between species and between genes. These differences are easily observable and measurable, as for example with Fop statistics which are the frequency of optimal codons (preferred codons in highly expressed genes). However, to fully understand how this bias arose and is maintained in a set of genes, we need a model that studies codon usage in an evolutionary approach. We develop a model, inspired from Yang and Nielsen [1], that distinguishes evolution of coding sequences at nucleotidic, codons and amino acids level. Our model explicitely describes both the mutational bias of nucleotides and the selection of preferred codons over other synonymous ones.

Evolutionary model Our evolutionary model is implemented in Bio++ [2] which is a set of C++ librairies for bioinformatics molecular evolution. Our model is based on a preference parameter for each codon, relative to its synonymous ones: $\Phi_{aa}(i)$ is for codon i that code for amino acid aa. This parameter corresponds to the relative codon frequency, within its amino acid, if selection for codon usage was the only selective pressure acting on sequences. We also consider Ψ_{aa_i} parameter which is the frequency of amino acid aa coded by codon i. Our sequence evolution model is such that equilibrium frequency of codon i, noted π_i^*, is

proportional to:

$$\pi_i^* \alpha \quad \underbrace{\pi_{i_1}\pi_{i_2}\pi_{i_3}}_{\substack{\text{nucleotides} \\ \text{equilibrium frequency}}} \quad . \quad \underbrace{\Psi_{aa_i}}_{\substack{\text{amino acid} \\ \text{preference}}} \quad . \quad \underbrace{\Phi_{aa}(i)}_{\substack{\text{codon} \\ \text{preference}}}$$

Parameter π_{i_p} is the equilibrium frequency of nucleotide i_p with $p \in [1, 3]$ the nucleotide position within codon. With no selection in our model, every codons must have the same equilibrium frequency which is: $\pi^* = \frac{1}{61}$ (note that in this model, we do not consider stop codons). We perform non-stationary runs enabling us to compute frequencies at the root of the tree. It helps us to depict how codon bias evolves and to understand how selective pressure occurs.

Results and Perspectives

We apply this model on three strains of $E.coli$ [4]: K12, CFT073 and 0157:H7. We have 3,353 genes clustered into concatenates of 100 genes, by increasing Fop (codon usage bias intensity). We study relationships between codon usage bias and others parameters. We obtain, as awaited, a negative correlation between $\omega = \frac{dN}{dS}$ and strong usage bias (high Fop). As expected, preferred codons (codons with highest $\Phi_{aa}(i)$) and tRNA content are positively correlated. GC content at equilibrium (GC*) is influenced by both selection on codon usage and mutational bias. More precisely, codons and nucleotides levels present contrary effects on base composition: codons level tends to enrich sequences with C and G whereas nucleotides level induced enrichment by A and T. With this non-stationary model, we can infer root and equilibrium frequencies of codons and we observe a global nucleotide enrichment towards AT.

We show there are opposite forces that drive sequence evolution from which selection on codon usage and also, mutational bias. We are currently refining the model at the amino acid level by considering distance between amino acids. We plan to use deeper datasets and non-homogeneous models of codon bias evolution.

References

[1] Yang, Z. and Nielsen, R: Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol 25(3):568-579, Mar (2008)

[2] Gueguen L., Gaillard S., Boussau B., Gouy M., Groussin M., Rochette NC., Bigot T., Fournier D., Pouyet F., Cahais V., Bernard A., Scornavacca C., Nabholz B., Haudry A., Dachary L., Galtier N., Belkhir K., Dutheil JY: Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol 30(8):1745-50, Aug (2013)

[3] Sharp, PM., Emery, LR. and Zeng K: Forces that influence the evolution of codon bias. Phil. Trans. R. Soc. B 365, 1203-1212 (2010)

[4] Jordan, IK, Kondrashov, FA, Adzhubei, IA, Wolf, YI, Koonin, EV, Kondrashov, AS and Sunyaev, A: A universal trend of amino acid gain and loss in protein evolution. Nature, 433(7026):633-638, Feb (2005)