CUBES and **CUBACR**, script packages for the analysis of evolutionary traits of codon bias.

Mauricio Javier Lozano*, José Luis López , María Laura Fabre, and Antonio Lagares

IBBM - Instituto de Biotecnología y Biología Molecular, CONICET, CCT-La Plata, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, calles 47 y 115, 1900-La Plata, Argentina..

Download a copy!

BACKGROUND

The balance between mutational biases and natural selection generates a wide range of GC contents and codon usage biases in Prokaryote genomes. Synonymous codons are selected to optimize translation of genes with different functions and expression levels, and to preserve the fitness of the cell. The choice of such optimal codons produce intragenomic codon usage heterogeneities[3]. The analysis of core-gene sets with increasing ancestries in bacteria, revealed an increased degree of adaptation of the most ancestral genes to the translational machinery[8]. This adaptation could be a consequence of codon selection for better translation accuracy or efficiency. One way to attempt differentiating those effects is to compare the codon usage of conserved vs. non-conserved genetic regions. Here we present bioinformatic tools which can perform several codon usage analysis on sets of genes belonging to different core-genomes.

What is a core-genome?

A core-genome is a set of genes which present an orthologous in all (or a high proportion) of the genomes in a given collection —i.e. all the common genes. Adding more genomes to the collection will reduce the core-genome, generally maintaining the genes which are related to fundamental cellular functions, and that were most likely present in the common ancestor. Using software such as EDGAR[2] or Get-Homologues[4], it is possible to search for all the orthologous genes between any number of genomes. Opposite to core-genomes, singletons are genes which are only present in one (or a low proportion) of the genomes under study, and are usually more related to newly acquired accessory functions.

What is the codon usage Bias?

Proteins are polymers of 20 different types of amino acids, each one encoded in the messenger RNA -which in turn is transcribed from the DNA- by a codon (a sequence of 3 nucleotidic bases). Since there are 4 types of nucleotides (A,T,C,G), a total of 64 codons exist. Of these, 3 codons (TAA, TTA, TGA) are used as translation stop signals, while the remaining 61 are used to encode for the 20 amino acids. The fact that there are more codons that amino acids makes the genetic code degenerate. Some amino acids are encoded by 1 codon only, while others are encoded by 2, 3, 4 or 6 codons. However, all the codons encoding the same amino acid -codon usage- are not used with the same frequency. This generates a codon usage bias which has been associated to gene-expression level, the adaptation to the proportion of transfer RNAs (tRNAs) of the cell, the degree of sequence conservation, the genomic location —i.e., chromosome, chromid, plasmidome[8]— and different features such as codon ramps, and mRNA secondary structure among others[3].

Measuring codon usage Bias

There are different kind of indices that quantify the codon bias of a gene. Some quantify the change in the frequency of use of each synonymous codon in relation to the frequency of random codon usage —e.g. Relative synonymous codon usage (RSCU, ENC)— while others compare this frequency to the codon usage bias observed on set of (predicted) highly expressed (PHE) proteins— e.g. Fop, CAI. Additionally, the translation adaptation index[6] (tAI) determines the degree of adaptation of the codon usage to the amount of each tRNAs of the cell[3].

And for a set of genes?

The obvious choice appears to be the average codon usage frequency of all the genes in the set. However, a method to calculate a frequency that represents in a more accurate way the codon usage frequency of the majority of the genes, denominated the modal frequency, was reported[5]. This method iteratively adjusts the codon usage frequency while comparing, by means of Chi-square statistics, to the codon usage frequency of all the genes in the set, until there is no statistically significant difference for most of the genes.

CUBES

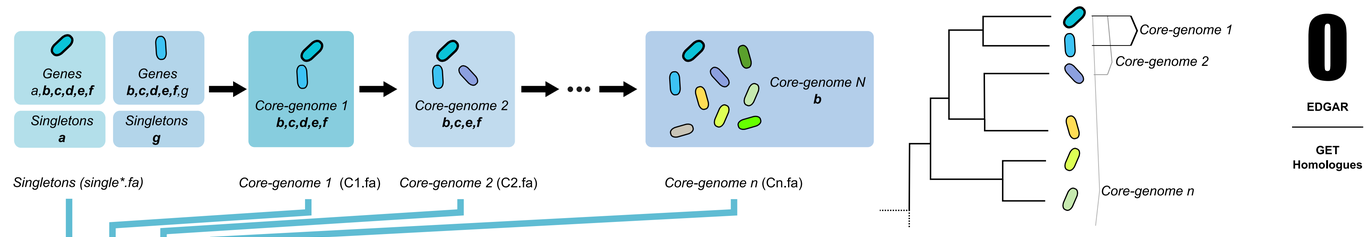
CUBES is a set of linux bash and perl scripts which can be used to:

1
- Calculate modal codon usage frequencies for sets of core-genomes (e.g. from a core-genome obtained from Edgar software)
- Generate a representative DNA sequences for those modal frequencies (Since codon usage analysis will be done using codonW[10], which requires coding sequences)
- Calculate the correspondence analysis (COA) of RSCU using CodonW software.
- Generate plots of the first 2 components of COA, both for genes and codons, including the modal sequences.

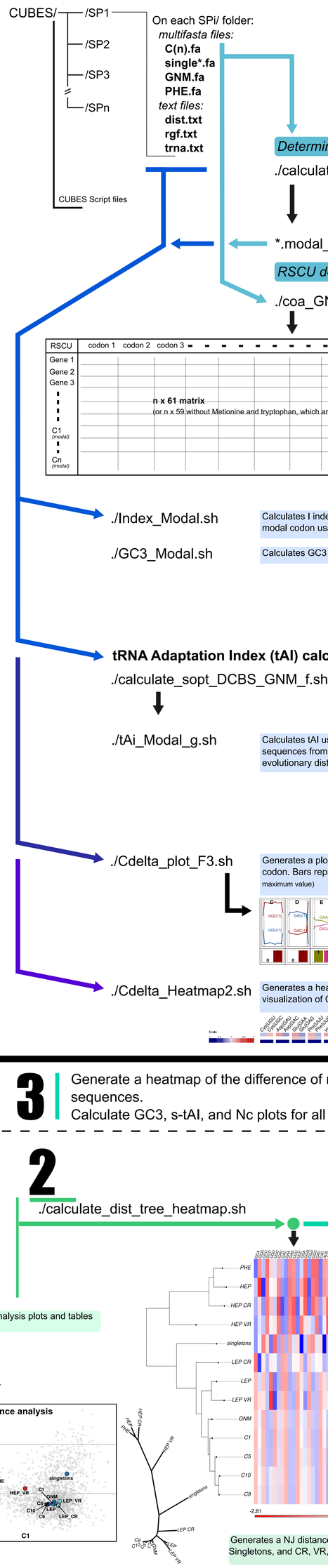
2 Calculate the variation of the adaptation index s-tAI and the GC3 content of the Core-genome Ci vs the evolutionary distance between the reference genome and the common ancestor of the core-genome Ci. Requires a trna.txt file with the tRNA genes copy number for the reference species, and the evolutionary distance between the most distant genomes in the core-genome. The script outputs tables and plots showing the variation of these indexes in function of the evolutionary distance.

3 Generate a plot showing the change in the codon usage frequencies (CUF) for each codon and the corresponding adaptiveness *w.*

4 Generate a heatmap of the tRNA adaptiveness (*w*), the difference in CUF between the C1 (initial core-genome) and the most ancestral Core-genome (Cn), and between putatively highly expressed genes (PHE) and Cn.



CUBES Pipeline



Core-genomes sets are constructed by the successive incorporation of new genomes to the core-genome analysis software (e.g. EDGAR[2] / Get_Homologues[4]), following the phylogeny of the family/ genus. Next, each of these Cn gene sets are used to calculate the relative synonymous codon useage (RSCU), the modal codon use frequency for each set and its RSCU, and to perform a Correspondence analysis of RSCU.

Required files description: multifasta files coding sequences in the complete genome, Ci, singletons, and PHE sets: C(n).fa, single*.fa, GNM.fa and PHE.fa. Additional text files: dist.txt (distance between the species, extracted from the species phylogenetic tree), rgf.txt (normalized frequency based on the number of tRNAs for each codon/amino-acid) and trna.txt (the number of copies of tRNA genes)

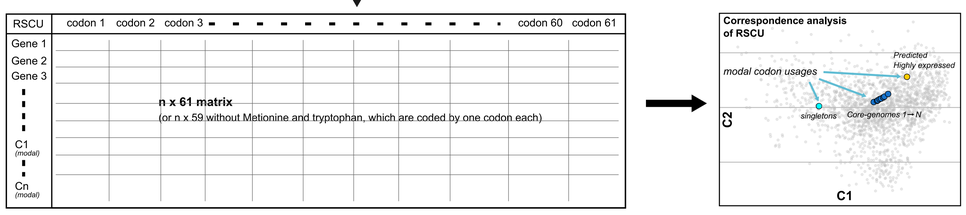
Determination of the modal codon usage bias

./calculate_modals2.sh — Scans all the SPi folders and calculates modal codon use frequencies, and generates a representative DNA sequence.

*.modal_freq.fas

RSCU determination and Correspondece analysis

./coa_GNM.sh — Calculates CA on RSCU, and generates several plots in R (Requires single*.fa, and GNM.fa)

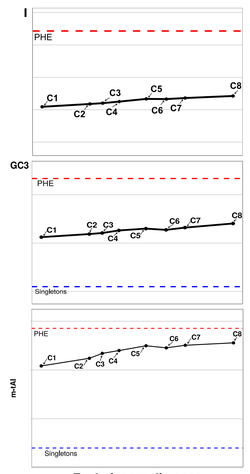


Correspondence analysis of RSCU **1**

./Index_Modal.sh — Calculates I index (described on Lopez JL. et al.) from modal codon usage frequencies.

./GC3_Modal.sh — Calculates GC3 index using modal codon usage.

All these scripts generate plots of calculated index vs evolutionary distance.



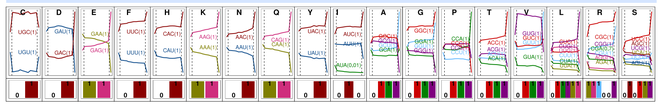
Calculate I,tAI or GC3 vs Evolutionary distance plots **2**

tRNA Adaptation Index (tAI) calculation

./calculate_sopt_DCBS_GNM_f.sh — Estimates Sij from the correlation between tAI and DCBS using the complete genome.

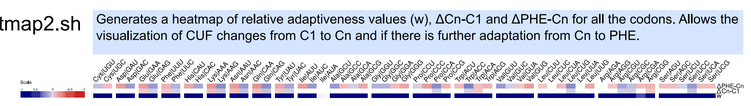
./tAi_Modal_g.sh — Calculates tAI using Sopt-DCBS-GNM for modal sequences from Ci->Cj, and generates a plot of tAI vs evolutionary distance.

./Cdelta_plot_F3.sh — Generates a plot showing the evolution of the frequency of synonymous codon usage from C1 to Cn for every codon. Bars represents the tRNA copy number. Between brackets is the *w* (calculated without the normalization by the maximum value)



Codon usage frequency over evolutionary distance **3**

./Cdelta_Heatmap2.sh — Generates a heatmap of relative adaptiveness values (w), ΔCn-C1 and ΔPHE-Cn for all the codons. Allows the visualization of CUF changes from C1 to Cn and if there is further adaptation from Cn to PHE.



Heatmap of w, ΔCn-C1 and ΔPHE-Cn **4**

CUBACR

Scripts for the codon usage bias analysis of Conserved and Variable regions from core genome proteins. CUBACR scripts are programmed to process multifasta files containing orthologs for each gene of a desired core-genome set (or sets of proteins with a defined expression level), make an amino acid guided codon alignment for each gene (using TranslatorX software[1]), and to output the conserved and variable regions of all genes of the set.

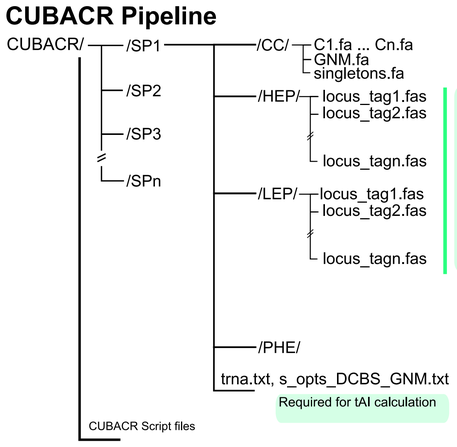
CUBACR can then be used to:

1
Make an amino acid based codon alignment and extract conserved and variable codons.
Calculate the modal codon usage frequencies for CR and VR.
Perform the correspondence analysis of RSCU for all genes and modal frequencies (Including CR and VR).

2 Calculate a neighbor joining tree using G. Olsen freqs_2_nj_tree_linux script[9] (Type 2 distances between CUF. Requires Phylip[7] software), and generate a heatmap of modal codon usage frequences guided by the previously generated tree.

3 Generate a heatmap of the difference of modal codon usage between Highly/lowly expressed and Conserved/Variable sequences.
Calculate GC3, s-tAI, and Nc plots for all the genes and modal sequences corresponding to each of the defined sets.

CUBACR Pipeline

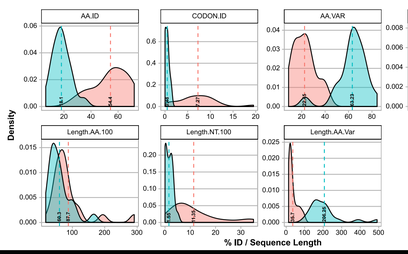


Multifasta files containing the sequences of the orthologs for locus_tagi in all the genomes included in a set.
(Highly / Lowly expressed proteins should be located on HEP and LEP respectively)

1

./cCvsE.sh

Sequences representing Modal CUFs
SPi_HEP_CR.modal_freq.fas
SPi_HEP_VR.modal_freq.fas
SPi_LEP_CR.modal_freq.fas
SPi_LEP_VR.modal_freq.fas
Ci.modal_freq.fas
GNM.modal_freq.fas
singletons.modal_freq.fas
CA_genes.svg
CA_codons.svg
Correspondence analysis plots and tables
SPi_CE.table
coverage_plot.svg
a table with the proportion of conserved and variable regions for each gene, and a density plot of the length of conserved and variable regions for HEP and LEP.



2

./calculate_dist_tree_heatmap.sh

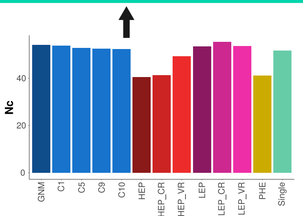
Generates a NJ distance tree and a heatmap of CUF for Ci, Singletons, and CR, VR, HEP and LEP sets.

3

./CRdelta_heatmap.sh

Calculates RSCU difference between HEP-CR, HEP and PHE, and plots different heatmaps. (Requires the outputs of Nc_plots.sh)

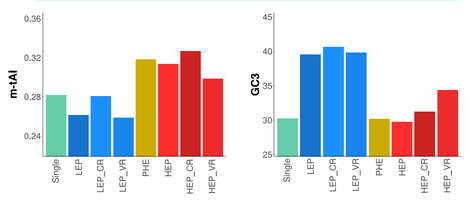
./Nc_plots.sh

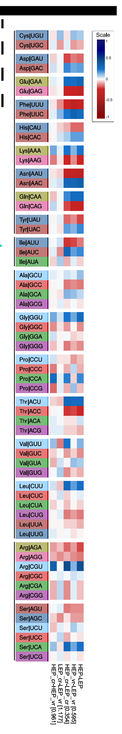


./GC3.sh

./tAi_Modal.sh

Requires:
trna.txt (relative frequency of tRNA gene copies)
s_opts_DCBS_GNM.txt (Sij obtained with CUBES using all the genome)





References

1. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38, 7–13 (2010).
2. Blom, J. et al. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res. 44, W22–W28 (2016).
3. Cannarozzi, G. & Schneider, A. Codon evolution - Mechanisms and Models. Oxford University Press (Oxford University Press, 2012). doi:10.1002/9783527678679.dg02069
4. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl. Environ. Microbiol. 79, 7696–7701 (2013).
5. Davis, J. J. & Olsen, G. J. Modal codon usage: Assessing the typical codon usage of a genome. Mol. Biol. Evol. 27, 800–810 (2010).
6. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res. 32, 5036–5044 (2004).
7. Felsenstein, J. PHYLIP (Phylogeny Inference Package). (2005).
8. López, J. L. et al. Codon Usage Heterogeneity in the Multipartite Prokaryote Genome: Selection-Based Coding Bias Associated with Gene Location, Expression Level, and Ancestry. MBio 10, 1–20 (2019).
9. Olsen, G. J. Codon Usage Utilities. (2010).
10. Peden, J. Correspondence Analysis of Codon Usage - CodonW. (2005).
11. Team, R. C. R: A language and environment for statistical computing. (2014).