

Comparison of synonymous substitution rates between paired- and unpaired nucleotides

Scripts used for this analysis were written in Python and R with objective to determine whether within a given protein coding alignment paired nucleotides have significantly lower synonymous substitution rates than unpaired nucleotides.

System and software requirement:

- This analysis requires a computer cluster and MPI libraries
- Python (<http://www.python.org/getit/>) must to be installed
- The HYPHY package should be installed on the system

Step 1: Running GARD, PARRIS and FUBAR

Source:

http://web.cbio.uct.ac.za/~brejnev/downloads/ComputationalTools/Scripts_used_to_run_GARD_PARRIS_and_FUBAR.zip

Put all the alignments in FASTA format within the “input” directory, change to the “bfiles” directory, run the “GARD_BF.py” script which will generate all required shell and batch files to run GARD. The output files in NEXUS format to be used as input for PARRIS and FUBAR will be written in the output folder. Run the PARRIS_BF.py and FUBAR_BF.py which will generate all shell and batch files required to run PARRIS and FUBAR. Run the shell scripts as recommended on your computer cluster. The final output will be written in the “output” folder.

N.B. the path to the location of programs must be changed within the python scripts.

Step 2: Comparing synonymous substitution rates between paired and unpaired nucleotides

Source:

http://web.cbio.uct.ac.za/~brejnev/downloads/ComputationalTools/Comparison_dS_rates_paired_and_unpaired_sites.zip

(1) Mapping the synonymous substitution rates to the alignments used by NASP for base-pairing prediction:

Run the script “Synonymous_substitutions-FUBAR.py” (in case you want to use rates obtained from FUBAR) or “Synonymous_substitutions-PARRIS.py” (in case you want to use rates obtained from PARRIS). For each of the datasets, the script maps the synonymous substitutions rates from the gene alignment to the full genome alignment used by NASP and produces a table with columns containing (i) the consensus sequence of the alignment used by NASP (ii) the consensus sequence of the gene

alignment – this is mapped to the NASP consensus sequence (3) a column containing either a “1” (indicating that a site is base-paired) or “0” (indicating that a site is not base-paired) at each row, and (4) the synonymous substitution rate for each site.

(2) Selecting rows corresponding to the third position in each codon within the alignments

Run the script “Devide_3rd_Positions-FUBAR.py” or “Devide_3rd_Positions-PARRIS.py” which will generate tables containing information only for third positions within codons.

(3) Mann-Whitney U test

To perform this test, run the “Non_Parametric_Test-FUBAR.py” or “Non_Parametric_Test-PARRIS.py” which will run “Non_Parametric_Tests_FUBAR.R” or “Non_Parametric_Tests_PARRIS.R” to perform a Mann-Whitney U test for every table generated in (2). The test gives statistical support on whether paired sites have lower synonymous substitution rates than unpaired sites.

Note: within the scripts the paths to the input files and executable programs must be updated.