HERMES v2.2

Guglielmo Puccio, Federico Plazzi

July 10, 2020

Contents

Chapter 1

License

Chapter 2

Introduction

2.1 Background: the tempo and mode of mitochondrial evolution

The HERMES index is a method of quantifying molecular evolution of mitochondrial genomes in different species and clusters; this method was originally proposed in [1]. The index relies on maximum likelihood factor analysis to summarize different measures that are typically found to be linked with evolutionary rates; it is intended to be computed *a posteriori*, i.e. after a phylogenetic and genomic analysis. As different empirical measures are merged together in a single score, it is a "hyper-empirical" index; moreover, it is a relative measure, because all species are compared with an outgroup: therefore, it was called *Hyper-Empirical Relative Mitochondrial Evolutionary Speed* (HERMES) index. The present Python script performs data collection and then calls a dedicated R [2] script to complete the factor analysis. The mitogenomic features that are currently implemented in HERMES-v2.2.py are:

1. the percentage of Unassigned Regions (URs);
2. the Amount of Mitochondrial Identical Gene Arrangements (AMIGA) index, which is defined as

$$\text{AMIGA} = \frac{N_{IGA} - 1}{N - 1} \tag{2.1}$$

   where N_{IGA} is the number of identical gene arrangements found in the dataset (taking only protein coding genes into account) with respect to a given species, and N is the total number of entries in the dataset;
3. the absolute value of the Strand Usage (SU) skew, which is defined as

$$\text{SUskew} = \frac{H - L}{H + L} \tag{2.2}$$

where H is the number of genes annotated on the leading strand and L is the number of genes annotated on the lagging strand;

4. the root-to-tip distance computed over a given phylogenetic tree;
5. the Maximum Likelihood (ML) pairwise distance from a given outgroup;
6. the AT content;
7. the AT skew, which is defined [3] as

$$\text{ATskew} = \frac{A - T}{A + T} \tag{2.3}$$

where A is the percentage of A in the genome and T is the percentage of T in the genome;

8. the GC skew, which is defined [3] as

$$\text{GCskew} = \frac{G - C}{G + C} \tag{2.4}$$

where G is the percentage of G in the genome and C is the percentage of C in the genome;

9. the number of (annotated) genes;
10. the length of the genome;
11. the Codon Adaptation Index, which is defined [4, 5] as

$$\text{CAI} = (\prod_{l=1}^{L} w_l)^{\frac{1}{L}} = (\prod_{l=1}^{L} \frac{N_{ij}}{N_{imax}})^{\frac{1}{L}} = e^{(\frac{1}{L} \sum_{l=1}^{L} \ln w_l)} \tag{2.5}$$

which is the geometric mean of the relative adaptiveness w of the L codons of the genome, excluding initiation and termination codons; in turn, the relative adaptiveness of the l-th codon of the genome (w_l) is defined as follows: if this codon is the j-th codon of a codon family for the amino acid i, its relative adaptiveness is its number of occurrences N_{ij} divided by the number of occurrences N_{imax} of the most frequently used codon (within its codon family) for the amino acid i – codons with no synonyms in a given genetic code would not contribute to the geometric mean and are therefore subtracted from L.

For each possible combination of at least two of these variables, a factor analysis is carried out. Normalization and varimax rotation are used, factor scores are found using correlation preserving, and correlations are found using the Pearson method; given the possible presence of missing values, missing data are set to be imputed using the median.
All the variables are pooled together for each species into the value of a single

loading: we define this score as the HERMES score of a given species. The best-performing variable set and the goodness-of-fit of the analysis is assessed following the recommendations of [6]: Tucker-Lewis Index (TLI; [7]) > 0.95; root mean square of the residuals (SRMR) < 0.08; root mean squared error of approximation (RMSEA) < 0.06; moreover, the Kaiser-Meyer-Olkin index (KMO; [8]) was taken into account on this regard.

2.2 What is HERMES.py for

HERMES.py is a Python script to

- compute some statistics that enter the HERMES index (namely, the UR proportion, the AMIGA index, the absolute value of the SU skew, the AT content, the AT skew, the GC skew, the number of genes, the length of the mtDNA, and the CAI.);
- compute the root-to-tip distance over a given phylogenetic tree;
- run RAxML [9] to compute the pairwise ML distance.

After this collecting phase, HERMES-v2.2.py calls an R script to perform the final factor analysis, whose result is the HERMES index itself.

2.3 What is HERMES.py not for

HERMES.py is not designed to produce all the requested data alone, because phylogeny is a complex task: many different tools and packages are available, and the user will choose the most suitable for her/his own needs. Therefore, HERMES-v2.2.py will not:

- produce a phylogenetic tree: a phylogenetic tree must be provided in Newick format;
- compute pairwise ML distance by itself: a pre-compiled RAxML distribution is provided to this purpose.

Eventually, the graphical parameters for the HERMES final plot are customizable only for what concerns colours: if more editing is needed, a table with raw data is provided as output, so that user can import it into a chart-editing software, or into R itself.

2.4 Citing HERMES

If you include HERMES-v2.2.py and/or the HERMES index in your publication, please cite [10]:

Plazzi F, Puccio G, Passamonti M. HERMES, or a Herald of the Mitochondria: an Improved Method to Test Mitochondrial Genome Molecular Synapomorphies among Clades. Genome Biol Evol. 2020;submitted.

As HERMES.py relies on R and RAxML for its analysis, you should also properly cite these softwares. Moreover, the package `ete` [11] is required by Python and the package `psych` [12] is required by R.

2.5 Who wrote HERMES.py

HERMES was written by Guglielmo Puccio and Federico Plazzi at the Department of Biological, Geological and Environmental Sciences of the University of Bologna. Thanks are also due to Elsa Spagnol for testing. Please report any bug to

`federico [dot] plazzi [at] unibo [dot] it`

Chapter 3

Running HERMES.py

3.1 Requirements

HERMES-v2.2.py is a Python script and calls an R script. To download and install Python and R, please refer to their sites:

- Python site
- R project site

Python requires the packages BioPython and `ete2`.
BioPython is generally already included in a typical Python distribution; for further information, the user is referred to the BioPython web site.
It is suggested to install `ete2` through the `pip` tool. In a Unix environment, if this is not available, just install it by typing

```
apt-get install python-pip
```

at the shell prompt (it requires root privileges). To install `ete2`, just type

```
pip install ete2
```

In a Windows environment, please refer to the `ete` web site for installation instructions.
A pre-compiled RAxML binary is provided with the HERMES package: this should work on most OSs. However, it is possible to use another version of RAxML by simply overwriting the provided `raxml` file with the preferred distribution. Please note that the RAxML executable must

- be in the HERMES folder;
- be called `raxml`.

The provided RAxML binary was originally compiled as `raxmlHPC-PTHREADS-SSE3` to enable multithreading.
The additional package `psych` is required by R in order to perform the factor analysis. To install it, just type

```
install.packages("psych")
```

at the R prompt and follow on-screen instructions. You may need root privileges.

3.2 Call the script

3.2.1 Basic syntax

The main HERMES-v2.2.py script is called through Python: the typical syntax is

```
python HERMES.py [options]
```

If no paths are provided, all the input files must be in the same directory where the HERMES package was placed. A child directory (`Results`) will be created with output files.
Input file formats (3.4) and all options (3.5) are detailed below.

3.2.2 Quick start

To test the package, just type

```
python HERMES-v2.2.py -I Sample/Annotations.gb -D Sample/Names.csv
-L Sample/Tree.tre -s Sample/Alignment.phy -m PROTGAMMAJTTF -q Sample/Partitions.txt
-O SoVe
```

from the command line (within the HERMES folder).

3.3 Package files

Four main files compose the HERMES package.

1. `HERMES-v2.2.py` is the main Python script.
2. The file `Genes.dict` is parsed by Python to set up a dictionary with gene names, which in turn is used to parse the GB file (see 3.4.1).
3. The file `colors` is parsed by Python to set up a dictionary with color names, which will be passed to R.

4. The R script is contained in the file `HERMES-vx.x`, where "`x.x`" is the version number.

3.4 Input files

Many different input files are requested or may be requested by HERMES.py; a sample file of each kind, taken from the analysis of [1], is provided along with the package files in the folder `Sample`.

3.4.1 GB file

This is either a GenBank-formatted file which contains the annotations of all the required complete mitochondrial genomes or a list of the relative Accession Numbers. In both cases, please note that the same Accession Numbers must be listed in the entry name file (see 3.4.2). If the extension of the GB file is `gb`, the Python script will look for a GenBank-formatted file. Annotated genes are taken from `CDS`, `rRNA`, and `tRNA` features, and only `CDS` features are used to compute the protein gene arrangement needed for the AMIGA index. There are many different ways to retrieve such a GenBank-formatted file: the easiest way is to locate a mitochondrial genome in the NCBI database and then click on "`Send:`" > "`File`", selecting "`GenBank`" as "`Format`". Many GenBank-formatted files can be concatenated in a single one using the `cat` command. For example, the command

```
cat *.gb > total.gb
```

will concatenate all `.gb` files in the `total.gb` file.
Otherwise, if many different entries are found, it is possible to follow the same pipeline as above to download a single GenBank-formatted file with all annotations.
A list file of NCBI Accession Numbers (one for each row) can be uploaded to Batch Entrez to retrieve all the requested entries together. Alternatively, il the extension of the GB file is `.list`, the Python script will read a list of Accession Numbers and fetch the annotations from GenBank as above.

3.4.2 Entry names and colours

A comma-separated table with three columns:

- The first column lists the name of OTUs both in the phylogenetic tree and in the alignment (see 3.4.3).
- The second column lists the corresponding GenBank Accession Numbers (see 3.4.1). If the Accession Number is lacking and this field is left blank, HERMES will look for custom annotations (see 3.5.12 and 3.5.13).

- The third column is optional and lists the colours that will be used for final plotting. This may be useful to create subsets of the dataset. Colours are listed through the R numeric notation, using numbers from 1 to 657.

3.4.3 Phylogenetic tree

A Newick-formatted phylogenetic tree. This will be used to compute root-to-tip distance.

3.4.4 Alignment file

A Phylip-formatted alignment file. This will be passed to RAxML to compute the pairwise ML distance, therefore it must be in compliance with RAxML standard (e.g., sequence names must be less than 256 characters in length; interleaved sequences are allowed; no spaces within sequences are allowed). The alignment must contain the outgroup.

3.4.5 Partition file

This is an optional file specifying partitions for the RAxML ML distance analysis. It must comply with RAxML requirements: please refer to the RAxML Manual for further details.

3.5 Options

Options are listed after the script call; their order is not important. Some options are mandatory, while other are optional flags: in the following, each option is discussed, specifying whether it is mandatory or optional .

3.5.1 `-h, --help`

Displays the help message and exits.

3.5.2 `-I`

`[mandatory]` Path to the GB file (see 3.4.1).

3.5.3 `-D`

`[mandatory]` Path to the entry name file (see 3.4.2). This should list correspondences between names in the tree and GB Accession Numbers, as well as colours/groupings.

3.5.4 `-L`

`[mandatory]` Path to the Newick tree file (see 3.4.3).

3.5.5 -O

[mandatory] Name of the outgroup, as written in the entry name file.

3.5.6 -G

[optional] The genetic code that will be applied to the whole dataset, following the GenBank numbering: currently, HERMES-v2.2 is not able to handle multiple genetic codes within the same dataset. It defaults to 5, the Invertebrate Mitochondrial Code.

3.5.7 -s

[mandatory] Path to the alignment file (see 3.4.4).

3.5.8 -m

[mandatory] ML model to be passed to RAxML, following RAxML syntax (e.g., "PROTCATJTTF", "GTRGAMMA"). Please refer to the -m option on the RAxML Manual.

3.5.9 -q

[optional] Path to the partition file (see 3.4.5).

3.5.10 -t

[optional] Number of threads to use to compute ML distances: this applies only if a multithreading version of RAxML is compiled (default). It defaults to 1.

3.5.11 -a

[optional] Level of significance requested for the RMSEA goodness-of-fit statistics. It defaults to 0.05.

3.5.12 -A

[optional] This flag allows to pass to HERMES newly annotated (or manually corrected) mitochondrial genomes, that are not available in GenBank. The annotation must be provided as a comma-separated table with four columns:

1. name of the gene (which must match an entry in the Genes.dict file);
2. strand (either "H" or "L");
3. start position;
4. end position.

In the row immediately above each table the entry name (as listed in the entry name file; see 3.4.2)must be placed; multiple tables can be concatenated in a single file.

3.5.13 `-F`

`[optional]` A FASTA file with the newly annotated (or manually corrected) mitochondrial genomes. The name of each genome must match the corresponding name in the entry name file (see 3.4.2).

Chapter 4

Output

4.1 Output files

4.1.1 HERMES plot

The HERMES plot is produced as `HERMES-v2.2_plot.pdf`, using colours if provided. Points are identified using entry names (as listed in the entry name file; see 3.4.2); please note that, in case of long names, these may fall outside the plotting region and the image may have to be edited further. Points are plotted in increasing HERMES order; exact HERMES values are provided in the `HERMES-v2.2_scores.txt` file (see 4.1.4).

4.1.2 Variables

Statistics for single genomes are listed in the file `HERMES_variables.txt` file. This file is produced at the end of the data collecting phase and is read by the R script to perform factor analysis.

4.1.3 Goodness-of-fit tests

Goodness-of-fit test results of the top-scoring combination are saved to the file `HERMES-v2.2_parameters.txt`. We suggest to read them following the recommendations of [6]: Tucker-Lewis Index (TLI; [7]) > 0.95; root mean square of the residuals (SRMR) < 0.08; root mean squared error of approximation (RMSEA) < 0.06. The Kaiser-Meyer-Olkin (KMO; [8]) measure of sampling adequacy is the proportion of variance that *might* be common variance: the higher the KMO index, the higher the reliability of the single loading selected by the factor analysis (i.e., the HERMES score). Finally, single communalities and the total communality are provided.

4.1.4 Scores

Single final HERMES scores are listed for each genome in the `HERMES-v2.2_scores.txt` file. This is provided in order to import it into another graphical software for further editing.

4.1.5 Best-fitting combinations and ties

The best-fitting combinations are listed in the `best.HERMES.out` file. If multiple combinations are present in this file, the first one is the one which was used for the final analysis and which is also reported in file `HERMES-v2.2_parameters.txt` (see 4.1.3). The HERMES tool identifies the best-fitting combination by counting how many goodness-of-fit tests are passed for each combination and adding the number of communalities >50% (the result is printed in column `parameters`). If a single combination maximizes this value, a single combination will be printed here.

In case of ties, multiple combinations will be printed here, and a "performance" of each combination is computed. In fact, each goodness-of-fit test yields a value: the ratio is computed between the threshold and the value (for statistics that have a minimum threshold – KMO, TLI, and communalities) or between the value and the threshold (for statistics that have a maximum threshold – SRMR and RMSEA). The best set of variables is the one that minimizes the mean of such ratios. Combinations are printed in decreasing order pf performance mean.

Bibliography

[1] F. Plazzi, G. Puccio, and M. Passamonti. Comparative large-scale mitogenomics evidence clade-specific evolutionary trends in mitochondrial DNAs of Bivalvia. *Genome Biol Evol.*, 8:2544–2564, 2016.

[2] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, 2013.

[3] A. Reyes, C. Gissi, G. Pesole, and C. Saccone. Asymmetrical directional-mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol.*, 15:957–966, 1998.

[4] P. M. Sharp and W. H. Li. The Codon Adaptation Index–a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications. *Nucleic Acids Res.*, 15:1281–1295, 1987.

[5] X. Xia. An Improved Implementation of Codon Adaptation Index. *Evol Bioinform.*, 3:53–58, 2007.

[6] L.-T. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6:1–55, 1999.

[7] L. R. Tucker and C. Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38:1–10, 1973.

[8] H. F. Kaiser. A second generation little jiffy. *Psychometrika*, 35:401–415, 1970.

[9] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313, 2014.

[10] F. Plazzi, G. Puccio, and M. Passamonti. HERMES, or a Herald of the Mitochondria: an Improved Method to Test Mitochondrial Genome Molecular Synapomorphies among Clades. *Genome Biol Evol.*, submitted, 2020.

[11] J. Huerta-Cepas, J. Dopazo, and T. Gabaldón. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, 11:24, 2010.

[12] W. Revelle. *psych: Procedures for Personality and Psychological Research.* Northwestern University, Evanston, 2014.