Geneplast: *R*/*Bioconductor* package for evolutionary rooting and plasticity inference based on distribution of orthologous groups.

Rodrigo JS Dalmolin*, Mauro AA Castro†

October 17, 2016

Contents

1 Overview

Geneplast is designed for evolutionary and plasticity analysis based on the distribuion of orthologous groups in a given species tree. It uses Shannon information theory to estimate the Evolutionary Plasticity Index (*EPI*) [**?**, **?**].

Figure ?? shows a toy example to illustrate the analysis. The observed itens in **Figure ??a** are distributed evenly among the different species (*i.e.* high diversity), while **Figure ??b** shows the opposite case. The diversity is given by the normalized Shannon's diversity and represents the distribution of orthologous and paralogous genes in a set of species. High diversity represents an homogeneous distribution among the evaluated species, while low diversity indicates that few species concentrate most of the observed orthologous genes.

The *EPI* characterizes the evolutionary history of a given orthologous group (OG). It accesses the distribution of orthologs and paralogs and is defined as,

$$EPI = 1 - \frac{H\alpha}{\sqrt{D\alpha}}, \tag{1}$$

where $D\alpha$ represents the OG abundance and $H\alpha$ the OG diversity. Low values of $D\alpha$ combined with high values for $H\alpha$ indicates an orthologous group of low plasticity, that is, few OG members distributed over many species. It also indicates that the OG might have experienced few modifications (*i.e.* duplication and deletion episodes) during the evolution. Note that $0 \leq H\alpha \leq 1$ and $D\alpha \geq 1$. As a result, $0 \leq EPI \leq 1$. For further information about the *EPI*, please see [**?**].

*rodrigo.dalmolin@imt.ufrn.br
†mauro.castro@ufpr.br

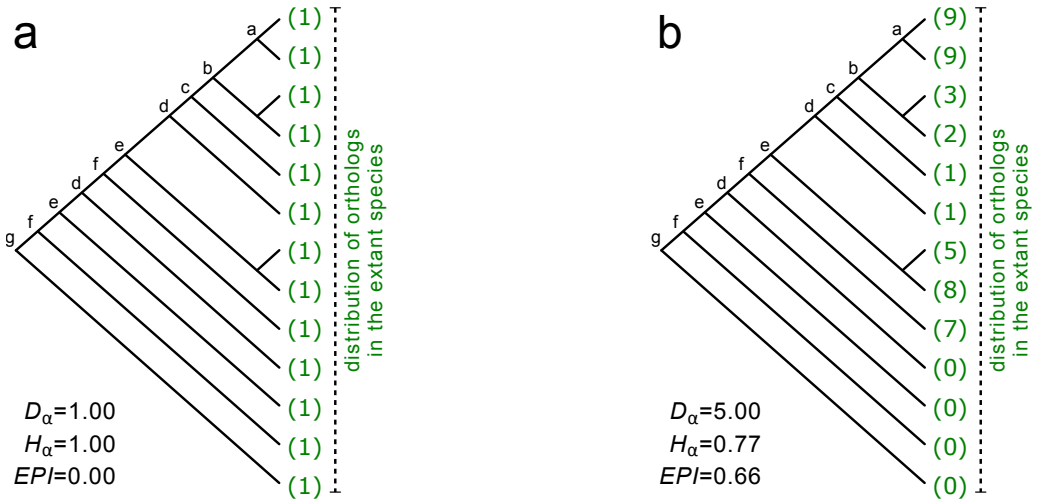

Figure 1: Toy examples showing the distribution of orthologous and paralogous genes in a given species tree. (**a**) OG of low abundance (Dα), high diversity (Hα) and consequently low plasticity (*EPI*). In this hypothetical case, the OG comprises orthologous genes observed in all species, without apparent deletion or duplication episodes. (**b**) in this example the OG is observed in many species, but not all, with many paralogs in some of them. Green numbers represents the number of orthologous genes in each species.

Geneplast also implements a new algorithm called *Bridge* in order to interrogate the evolutionary root of a given gene based on the distribution of its orthologs. The *Bridge* algorithm assesses the probability that an ortholog of a given gene is present in each last common ancestor (LCA) of a given species (in a given species tree). As a result, this approach infers the evolutionary root representing the gene emergence. The method is designed to deal with large scale queries in order to interrogate, for example, all genes annotated in a network (please refer to [**?**] for a case study illustrating the advantages of using this approach).

To illustrate the rooting inference consider the evolutionary scenarios presented in **Figure ??** for the same hypothetical OGs. These OGs comprise a number of orthologous genes distributed among 13 species, and the pattern of presence or absence is indicated by green and grey colours, respectively. Observe that at least one ortholog is present in all extant species in **Figure ??a**. To explain this common genetic trait, one possible evolutionary scenario could assume that the ortholog was present in the LCA of all species and was genetically transmitted up to the descendants. For this case, the evolutionary root might be placed at the bottom of the species tree (*i.e.* node *g*). The same reasoning can be done in **Figure ??b**, but with the evolutionary root placed at node *d*. The *Geneplast* rooting pipeline is designed to infer the most consistent rooting scenario for the observed orthologs in a given species tree. The pipeline provides a consistency score called *Dscore* which estimates the stability of the inferred root, as well as an associated empirical p-value computed by permutation analysis.

2 Quick start

The orthology data required to run *Geneplast* is available in the `gpdata.gs` dataset. This dataset includes four objects containing information about Clusters of Orthologous Groups derived from the STRING database, release 9.1. *Geneplast* can also be used with other sources of orthology information, provided that the input is set according to the `gpdata.gs` data structure (*note: in order to reduce the processing time this example uses a subset of the STRING database*).

```
> library(geneplast)
> data(gpdata.gs)
```

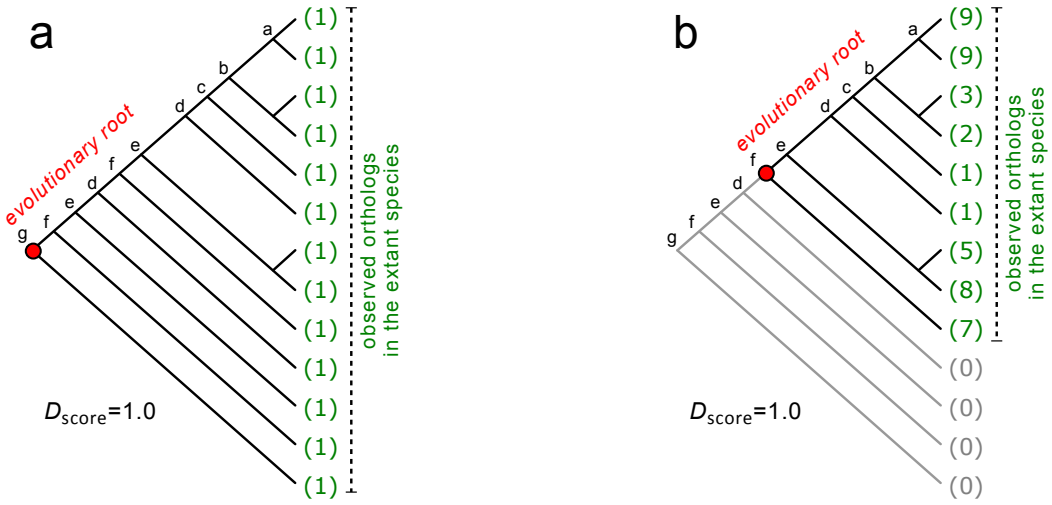

Figure 2: Possible evolutionary rooting scenarios for the same toy examples depitected in Figure 1. (**a, b**) Red circles indicate the evolutionary roots that best explain the observed orthologs in this species tree.

2.1 Evolutionary plasticity inference

The first step is to create an *OGP* object by running the `gplast.preprocess` function. This example uses 121 eukaryotic species from the STRING database and all *OGs* mapped to the genome stabilty gene network [**?**]. Next, the `gplast` function perform the plasticity analysis and the `gplast.get` returns the results:

- 1 - Create an object of class *OGP*.
```
> ogp <- gplast.preprocess(cogdata=cogdata, sspids=sspids, cogids=cogids, verbose=FALSE)
```

- 2 - Run the `gplast` function.
```
> ogp <- gplast(ogp, verbose=FALSE)
```

- 3 - Get results.
```
> res <- gplast.get(ogp,what="results")
> head(res)
        abundance diversity plasticity
KOG0011    1.7328    0.9532     0.2759
KOG0028    3.1466    0.9207     0.4809
KOG0034    4.1121    0.9216     0.5455
KOG0037    2.8252    0.9116     0.4577
KOG0045    7.3534    0.8965     0.6694
KOG0192   26.9286    0.8284     0.8404
```

The results are returned in a 3-column `data.frame` with OG ids (cogids) identified in `row.names`. Columns are named as *abundance*, *diversity*, and *plasticity*.

The metric *abundance* simply indicates the ratio of orthologs and paralogs by species. For example, KOG0011 cromprises 201 genes distributed in 116 eukaryotic species, with a resulting abundance of 1.7328. Abundance of 1 indicates an one-to-one orthology relationship, while high abundance denotes many duplication episodes on the OG's evolutionary history. Diversity is obtained applying normalized Shannon entropy on orthologous distribution and Plasticity is obtained by EPI index, as described equation (1).

2.2 Evolutionary rooting inference

The rooting analysis starts with an *OGR* object by running the `groot.preprocess` function. This example uses all *OGs* mapped to the genome stability gene network using *H. sapiens* as reference species[**?**] and is set to perform 100

permutations for demonstration purposes (for a full analysis, please set `nPermutations`$\geq$1000). Next, the `groot` function performs the rooting analysis and the results are retrieved by `groot.get`, which returns a `data.frame` listing the root of each OG evaluated by the `groot` method. The pipeline also returns the inconsistency score, which estimates the stability of the rooting analysis, as well as the associated empirical p-value. Additionally, the `groot.plot` function allows the visualization of the inferred root for a given OG (*e.g.* **Figure ??**) and the LCAs for the reference species (**Figure ??**).

- 1 - Create an object of class *OGR*.

```
> ogr <- groot.preprocess(cogdata=cogdata, phyloTree=phyloTree, spid="9606",
+                         cogids=cogids, verbose=FALSE)
```

- 2 - Run the groot function.

```
> set.seed(1)
> ogr <- groot(ogr, nPermutations=100, verbose=FALSE)
```

- 3 - Get results.

```
> res <- groot.get(ogr,what="results")
> head(res)
          Root Dscore   Pvalue AdjPvalue
NOG251516    3   0.67 2.49e-10  3.54e-08
NOG80202     4   1.00 1.46e-09  2.07e-07
NOG72146     6   0.82 2.99e-05  4.24e-03
NOG44788     6   0.56 1.61e-04  2.28e-02
NOG39906     7   1.00 8.30e-09  1.18e-06
NOG45364     9   0.83 1.94e-07  2.76e-05
```

- 4 - Check the inferred root of a given OG

```
> groot.plot(ogr,whichOG="NOG40170")
PDF file 'gproot_NOG40170_9606LCAs.pdf' has been generated!
```

- 5 - Visualization of the LCAs for the reference species in the analysis (i.e. *H. sapiens*)

```
> groot.plot(ogr,plot.lcas = TRUE)
PDF file 'gproot_9606LCAs.pdf' has been generated!
```

3 Session info

```
> toLatex(sessionInfo())
```

- R version 3.3.1 (2016-06-21), `x86_64-apple-darwin13.4.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: geneplast 1.0.0
- Loaded via a namespace (and not attached): BiocStyle 2.2.0, ape 3.5, grid 3.3.1, lattice 0.20-34, nlme 3.1-128, snow 0.4-2, tools 3.3.1

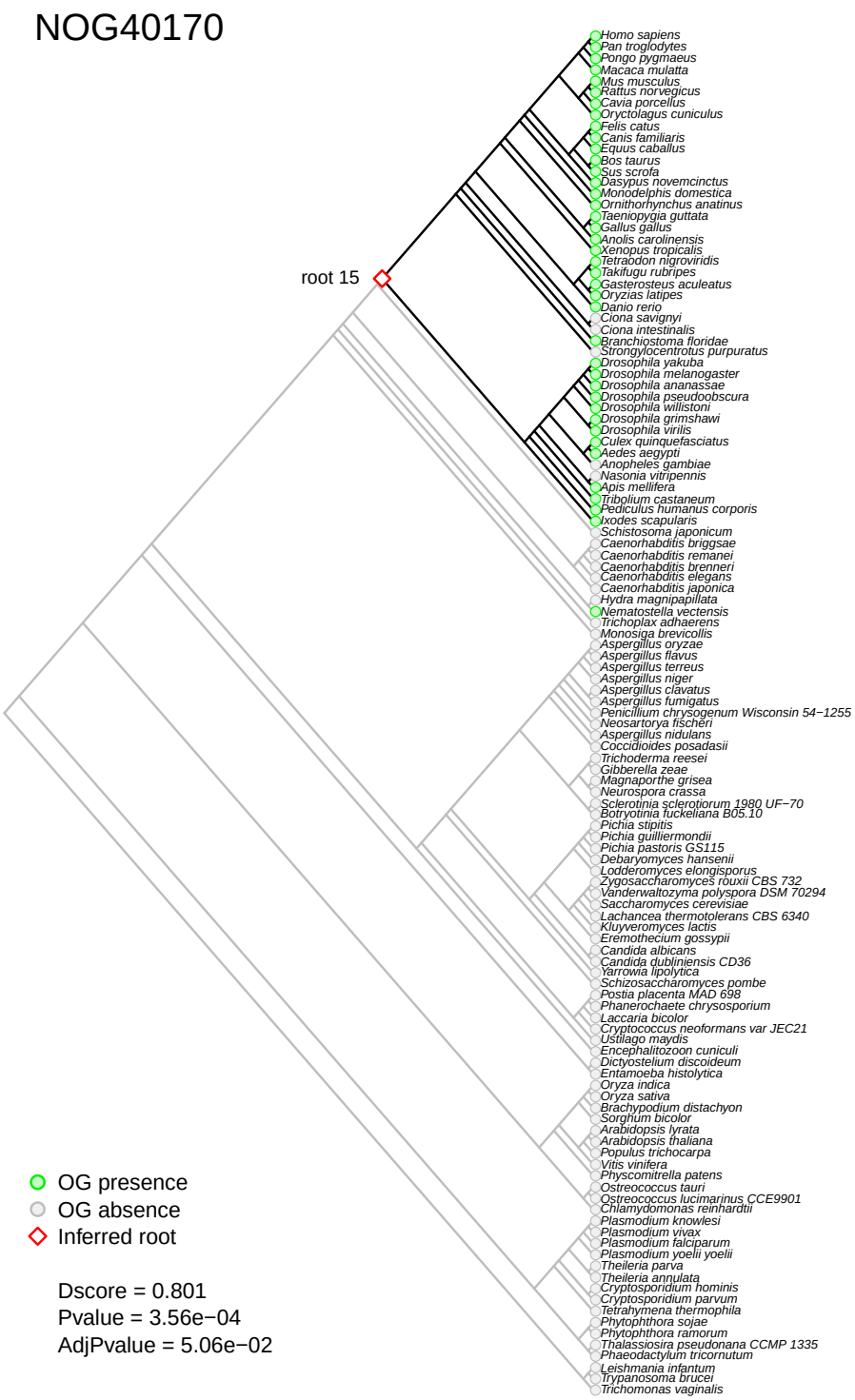

Figure 3: Inferred evolutionary rooting scenario for NOG40170. Monophyletic groups are ordered to show all branches of the tree below the queried species in the analysis.

Figure 4: Visualization of the LCAs for the reference species in the analysis.