User Manual for *CDROM* Source Code

Brent R. Perry and Raquel Assis

April 25, 2016

Overview

This manual provides step-by-step instructions for running *CDROM* from source. The *CDROM* R package and user manual for running *CDROM* as an R package are available on CRAN.

Introduction

CDROM is an R implementation of Assis and Bachtrog's (2013) method for classifying duplicate gene retention mechanisms via phylogenetic comparisons of gene expression data in two species. *CDROM* takes as input a table of duplicate genes in one species and their ancestral genes in a second species, a table of orthologous single-copy genes in the two species, and tables containing gene expression data for each species. First, *CDROM* obtains gene expression profiles by converting raw expression levels to relative expression values (proportions of contributions to total gene expression). Next, it computes Euclidian distances between the expression profiles of each duplicate gene and the ancestral gene ($E_{\mathrm{D1,A}}$ and $E_{\mathrm{D2,A}}$ by default, or $E_{\mathrm{P,A}}$ and $E_{\mathrm{C,A}}$ if parent/child copies are specified), the combined duplicate gene expression profile and the ancestral gene expression profile ($E_{\mathrm{D1+D2,A}}$ or $E_{\mathrm{P+C,A}}$), and the expression profiles of orthologous single-copy genes ($E_{\mathrm{S1,S2}}$). Then, it obtains a cutoff for expression divergence, E_{div}, which by default is set to the semi-interquartile range (SIQR) from the median of $E_{\mathrm{S1,S2}}$, but can also be user-defined. Last, it uses the following phylogenetic rules to classify retention mechanisms of duplicate genes: conservation if $E_{\mathrm{P,A}} \leq E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} \leq E_{\mathrm{div}}$; neofunctionalization if $E_{\mathrm{P,A}} > E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} \leq E_{\mathrm{div}}$, or if $E_{\mathrm{P,A}} \leq E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} > E_{\mathrm{div}}$; subfunctionalization if $E_{\mathrm{P,A}} > E_{\mathrm{div}}$, $E_{\mathrm{C,A}} > E_{\mathrm{div}}$, and $E_{\mathrm{P+C,A}} \leq E_{\mathrm{div}}$; or specialization if $E_{\mathrm{P,A}} > E_{\mathrm{div}}$, $E_{\mathrm{C,A}} > E_{\mathrm{div}}$, and $E_{\mathrm{P+C,A}} > E_{\mathrm{div}}$.

Setup

Navigate to www.personal.psu.edu/rua15/software.html. Download the CDROM.zip file, and extract it into your working directory. This file contains the CDROM R source code (`CDROM.R`), as well as four sample input files:

1) `human_chicken_dups`
   Parent, child, and ancestral genes for duplications that occurred after human-chicken divergence
2) `human_chicken_singles`
   Orthologous single-copy genes in human and chicken
3) `human_expr`
   Human gene expression levels in ten tissues
4) `chicken_expr`
   Chicken gene expression levels in ten tissues (same tissues as human)

In the remainder of this manual, we will demonstrate how to run *CDROM* from source with the sample input files provided, though it can be applied to any similarly-formatted dataset.

Running *CDROM* from source with the sample input files

In R, enter the following command to load `CDROM.R`:

```
>source("CDROM.R")
```

Next, type the following command to run *CDROM* with default parameters:

```
>CDROM(dupFile="human_chicken_dups", singleFile="human_chicken_singles",
exprFile1="human_expr", exprFile2="chicken_expr")
```

CDROM will take a few seconds to run, after which it will output three files:

1) `out.png`
   Figure showing distributions of all Euclidian distances calculated, as well as the location of E_{div}
2) `out1.txt`
   Text file providing classifications of retention mechanisms for genes in `human_chicken_dups`
3) `out2.txt`
   Text file providing counts of classifications obtained with five E_{div} values

Note that `out.png` should look like Figure 1A in the manuscript.

Parameters

CDROM has several parameters that can be modified:

`out`	The prefix to be used in the names of the two output files
`PC`	A logical value indicating whether parent and child copies are separated (defaults to `FALSE`)
`Ediv`	The divergence cutoff to be used in classifications (defaults to SIQR)
`useAbsExpr`	A logical value indicating whether absolute expression levels are used (defaults to `FALSE`)
`head1`	A logical value indicating whether `exprFile1` contains a header line (defaults to `TRUE`)
`head2`	A logical value indicating whether `exprFile2` contains a header line (defaults to `TRUE`)
`head3`	A logical value indicating whether `dupFile` contains a header line (defaults to `TRUE`)
`head4`	A logical value indicating whether `singleFile` contains a header line (defaults to `TRUE`)
`legend`	A keyword indicating the position of the legend in the output figure (defaults to `'topleft'`)

For example, to generate Figure 1B in the manuscript, type the following command:

```
>CDROM(dupFile="human_chicken_dups", singleFile="human_chicken_singles",
exprFile1="human_expr",exprFile2="chicken_expr", PC = TRUE)
```

Two parameters that may be of interest are `Ediv` and `useAbsExpr`. Modifying `Ediv` may aid in data exploration and assessing the robustness of classifications. Setting `useAbsExpr=TRUE` results in the calculation of Euclidian distances from absolute, rather than relative, expression levels. This is not recommended in most cases, though it is necessary when there is only data from one sample.

References

Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA.* 110:17409-17414 (2013).

Perry, BR, Assis R. CDROM: Classification of Duplicate gene RetentiOn Mechanisms. *BMC Evol. Biol.* DOI:10.1186/s12862-016-0644-x (2016).