Tutorial: Use *TreeExp* for Phylogenetic Transcriptome Analysis

Jingwen Yang

2019-05-29

Contents

Chapter 1

Introduction

1.1 Scope

This guide provides an brief overview of the package *TreeExp* which is developed to provides useful phylogenetic tools applicable to RNA-seq data.

Statistical methods implemented in the package was based on Ornstein-Uhlenbeck (OU) model of transcriptome evolution which claims that expression changes are constrained by stabilizing selection.

The package can be applied to comparative expression evolution analysis based on RNA-seq data, which includes but not liminited to:

- pairwise expression distance estimation
- relative rate test for transcriptome evolution
- the strength of expression conservation estimation
- ancestral transcriptome inference

This guide begins with brief description of the input data manipulation and storage and then gives key capabilities of package. Each main feature of the package consists of two parts: biological model and fully worked case studies for real data.

1.2 Installation

A convenient way to install package from github is through *devtools* package:

```
install.packages('devtools')
devtools::install_github("jingwyang/TreeExp")
```

After installation, *TreeExp* can be loaded in the usual way:

```
library('TreeExp')
```

1.3 Citation

The *TreeExp* package implements statistical methods from the following publications. If you use *TreeExp* in publised research, please cite the appropriate articles

```
Ruan,H. et al. (2016) TreeExp1.0: R package for analyzing expression
evolution based on rna-seq data. Journal of Experimental Zoology Part B:
Molecular and Developmental Evolution, 326, 394-402.
```

- This paper (Ruan et al., 2016) released the 1.0 version of *TreeExp* that can perform comparative expression evolution analysis based on RNA-seq data, which include optimized input formatting, normalizeation, pairwise expression distance estimation, expression character tree inference, and prliminary expression phylogenetic network analysis.

```
Yang,J. et al. (2018) Ancestral transcriptome inference based on
rna-seq and chip-seq data. Methods.
```

- This paper(Yang et al., 2018) reported an updated verson of ancestral state inference originally developed by (Gu, 2004). With special reference to the transcriptome evolution, the algorithm implemented is feasible, which can deal with RNA-seq and ChIP-seq data.

```
Gu, Xun, Hang Ruan, and Jingwen Yang. 2019. "Estimating the Strength of
Expression Conservation from High Throughput RNA-seq Data." Bioinformatics, May.
```

- This paper (Gu et al., 2019) developed a gamma distribution model to describe how the strength of expression conservation (denoted by W) varies among genes. Given the high throughput RNA-seq datasets from multiple species, we have formulated an empirical Bayesian procedure to estimate W for each gene.

1.4 How to Get Help?

Each function in *TreeExp* has online help page. If users have a question about a particular function, reading the function's help page will be very useful. For example, a detailed description of the arguments and output of the *RelaRate.test* function can be read by typing

```
?RelaRate.test()
```

or

```
help(RelaRate.test)
```

at R console. Users can also read the tutorial file embeded in the package to have more detailed information about the pacakge.

It seems that vigenette file was not build by default when installing the pacakge from github through *devtools* package.Before we can check the vigenettes, we should build it first.

One way is to build it when we install the package:

```
devtools::install_github('jingwyang/TreeExp',build_opts = c("--no-resave-data", "--no-manual"))
```

Then we can list available vigenettes in an HTML browser through `browseVignettes` function:

```
browseVignettes('TreeExp')
```

Besides, authors are appreciated to recieve reports of bugs in the functions or well-considered suggestions for improvements for the package.

1.5 RNA-seq Data Enbeded

RNA-seq datasets used in cased studies (brain, cerebellum, heart, liver, kidney and testis) were collected from the work of (Brawand et al., 2011), each of which include eigth species: Human (*Homo sapiens*), Chimpanzee (*Pan troglodytes*), Orangutan (*Pongo abelii*), Macaque (*Macaca mulatta*), Mouse (*Mus musculus*), Platypus (*Ornithorhynchus anatinus*), Opossum (*Monodelphis domestica*) and Chicken (*Gallus gallus*).

Single expression value for each gene per species was obtained by taking the median of TPM (Transcripts per Million) among biological replicates.

Chapter 2

Data Manipulation and Storage

2.1 Input Format

To simplify the use of package, the updated *TreeExp* has gave up the former input format (which require both gene information file and reads count data) and takes in only normalized RNA-seq data as input file. In other words, users should make sure that the input data is processed and comparable between samples. The package is not likely to provied data filteration or normalization functions.

The expression data is suppossed to be in certain format:

1. expression file should be a `text` file in matrix shape, in which values are separated by tabs. `Rows` correspond to orthologous genes names, and `Columns` correspond to sample names. Sample names are in format of "`TaxaName_SubtaxaName_ReplicatesName`". Usually, `TaxaName` represents name of species; `Subtaxaname` correspond to cetain tissue, celltype or develomental stage; `ReplicatesName` shows the name of replicates for each `Taxa_Subtaxa`-pair. The three types of lables, `TaxaName`,`SubtaxaName`, `ReplicatesName` are connected by '`_`' character.

2. raw reads count data should be first normalized, e.g., by RPKM. While RPKM is simple and straightforward, it tends to be unstable when the number of genes expressed across samples differs considerably. This problem can be alleviated by the TPM measure, which has been widely usued. Some statistically sophisticated normalization methods such as TMM, and median ratio normalization, has become the built-in standard in many bioinformatics tools for RNA-seq analysis(Robinson et al., 2010).

2.2 Example of Input Data

The example file are included in the *TreeExp* package, which can be found in `/inst/extdata` folder in the package.

Here, we select expression values of 100 orthologous genes in eight brain regions (CB, HIP, STR, ACC, V1C, PMC, DPFC, VPFC)[1] among human, chimpanzee, gorilla and gibbon(Xu et al., 2018). The numbers of biological replicates for each of the brain regions in species are 2~6, except only one replicate for all brain regions in gibbon. Note that expression data here are only used as demonstration of how functions in package store, manipulate and print the data input, and should not be used in further phylogenetic analysis since too few genes were included in the file.

The Table below shows the format of the partly input data.

[1]cerebellum (CB), hippocampus (HIP), striatum (STR) , anterior cingulate cortex (ACC), primary visual cortex (V1C), premotor cortex (PMC), dorsolateral prefrontal cortex (DPFC), ventrolateral prefrontal cortex (VPFC).

Gene	Human_DPFC_Hs3	Human_STR_Hs8	Chimpanzee_ACC_REIKO	Gorilla_CB_GON
ENSG00000000003	10.5	1.8	3.7	0.9
ENSG00000000005	0.0	0.0	0.0	0.0
ENSG00000000419	33.7	34.1	17.3	19.9
ENSG00000000457	1.3	1.9	1.2	3.7
ENSG00000000460	0.5	0.7	0.6	1.3

2.3 Construction

The construction function `TEconstruct` loads in expression level file, and wraps it in a list of *taxonExp* objects (one *taxaExp* object).

```
taxa.objects = TEconstruct(ExpValueFP = system.file('extdata/primate_brain_expvalues.txt',
package = 'TreeExp'), taxa = "all", subtaxa = 'all')
```

The construction process takes **several minutes** on a desktop computer depending on data size and hardware performance. Specify **"taxa"** and **"subtaxa"** options in the function when using partial of your data. The construction process will be faster.

```
taxa.objects = TEconstruct(ExpValueFP = system.file('extdata/primate_brain_expvalues.txt',
package = 'TreeExp'), taxa = "all", subtaxa = c("ACC","CB"))
```

You can take a look at what the loaded objects:

```
print(taxa.objects, details = TRUE)
```

```
##
##  8 taxonExp objects
##
## object 1 : Human      ACC
## object 2 : Human      CB
## object 3 : Chimpanzee     ACC
## object 4 : Chimpanzee     CB
## object 5 : Gorilla    ACC
## object 6 : Gorilla    CB
## object 7 : Gibbon     ACC
## object 8 : Gibbon     CB
```

Also, you can choose to print a single *taxonExp* object

```
print(taxa.objects[[1]], printlen = 6)
```

```
##
## One taxonExp object
## Taxon name:  Human
## Subtaxon name:  ACC
## Total gene number:  200
## Total bio replicates number:  5
## Bio replicates titles:
## [1] "Hs1" "Hs6" "Hs8" "Hs2" "Hs5"
```

and choose to print single element (*exp_val*(expression values) element as example) for the *taxonExp* object

```
taxa.objects[[6]]$exp_value[1:5,]
```

	Gorilla_CB_Sakura	Gorilla_CB_GON
ENSG00000000003	2.6	0.9
ENSG00000000005	0.0	0.0
ENSG00000000419	13.5	19.9
ENSG00000000457	4.1	3.7
ENSG00000000460	1.3	1.3

Once the contruction courcs successfully completed, the following transcriptome phylogenetic analysis are ready to go.

Chapter 3

Pairwise Expression Distance and Expression Tree-Making

3.1 Theory

3.1.1 The Ornstein-Uhlenbeck (OU) model of transcriptome evolution

It is generally believed that the expression level of a gene, denoted by x, is subject to the stabilizing selection to maintain the optimum during the course of evolution. Consequently, the motion of the expression level (x) can be described by the so-called Ornstein-Uhlenbeck (OU) process(Hansen and Martins, 1996, Bergmann et al. (2004)Butler and King (2004)). Briefing speaking, while the random force driving the expression level away from the optimum, the deterministic force will pull the expression level back to the optimum. It has been shown that the distribution density of x after t time units since the initial value x_0, denoted by $OU\left(x \mid x_0, \theta\right)$, is normal with the mean $E\left[x \mid x_0\right]$and variance $V\left(x \mid x_0\right)$ given by

$$\begin{aligned} E\left[x \mid x_0\right] &= \mu\left(1-e^{-\beta t}\right) + z_0 e^{-\beta t} \\ V\left[x \mid x_0\right] &= \frac{1-e^{-2\beta t}}{W} \end{aligned} \tag{1.1}$$

respectively, where θ is for the parameter vector: μ is the optimal expression value, β the rate of expression evolution, and W the strength of expression conservation.

3.1.2 Transcriptome evolution between species

Consider a simple two-stage scenario of transcriptome evolution that consists of two species that diverged t time units ago (**Figure 1A**).

The first stage is the evolutionary lineage from the tissue ancestor (denoted by node Z) to the common ancestor of two species (denoted by node O), with a span of τ evolutionary time units. Given the initial expression value z_0 at node Z, the density of x_0 (the ancestral expression level at node O) is given by $OU\left(x \mid x_0,\ \theta_0\right)$ where the parameter vector $\theta_0 = (\mu,\ \beta\tau,\ W)$.

In the second stage, let x_1 and x_2 be the expression levels of an orthologous gene pair, respectively. Given the ancestral expression level (x_0) at node O, the density of x_1 follows $OU\left(x_1 \mid x_0;\ \theta_1\right)$, and that of x_2 follows $OU\left(x_2 \mid x_0;\ \theta_2\right)$, respectively, where $\theta_1 = (\mu,\ \beta_1 t,\ W_1)$ and $\theta_2 = (\mu,\ \beta_2 t,\ W_2)$. If transcriptome evolution is independent between lineages, the joint density of x_1 and x_2 conditional of x_0 is simply given by

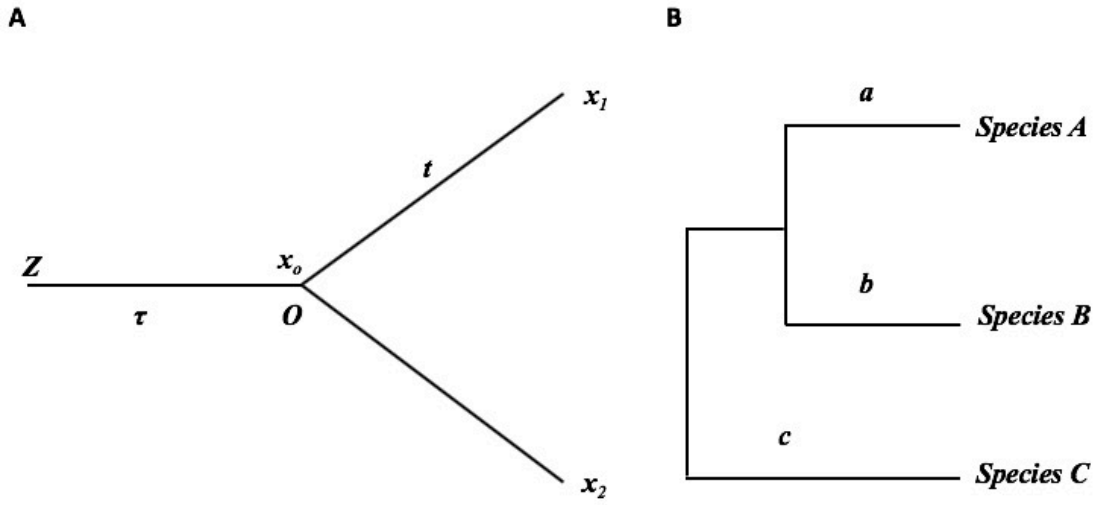


Figure 3.1: Figure1. The schematic of transcriptome evolution along a phylogeny

$P(x_1,\ x_2\ \mid\ x_0) = OU(x_1 \mid x_0;\ \theta_1)\, OU(x_2 \mid\ x_0;\ \theta_2)$. Since x_0 follows the density $\pi(x_0) = OU(x_0 \mid z_0;\ \theta_0)$, one can show that the joint density of x_1 and x_2

$$P(x_1, x_2) = \int_{-\infty}^{\infty} P(x_1, x_2 \mid x_0)\, \pi(x_0)\, dx_0 \tag{1.2}$$

follows a normal distribution, with the mean vector

$$\begin{aligned} E[x_1] &= \mu\left(1 - e^{\beta_1 t}\right) + \left[\mu\left(1 - e^{\beta_0 \tau}\right) + z_0 e^{\beta_0 \tau}\right] e^{-\beta_1 t} \\ E[x_2] &= \mu\left(1 - e^{\beta_2 t}\right) + \left[\mu\left(1 - e^{\beta_0 \tau}\right) + z_0 e^{\beta_0 \tau}\right] e^{-\beta_2 t} \end{aligned} \tag{1.3}$$

and the variance-covariance matrix $\boldsymbol{V}$given by

$$\begin{aligned} V_{11} &= \frac{1}{W_1} + \left(\rho^2 - \frac{1}{W_1}\right) e^{-\beta_1 t} \\ V_{22} &= \frac{1}{W_2} + \left(\rho^2 - \frac{1}{W_2}\right) e^{-\beta_2 t} \\ V_{12} &= \rho^2 e^{-(\beta_1 + \beta_2) t} \end{aligned} \tag{1.4}$$

where $\rho^2 = V(x_0 \mid\ z) = \frac{\left(1 - e^{-\beta\tau}\right)}{W}$ is the variance of x_0 at root O.

3.1.3 The estimation problem of expression distance

Consider two RNA-seq datasets of the same tissue from species-1 and species-2, respectively, which include N orthologous genes. For the k-th orthologous pair, let $x_{k,1}$ and $x_{k,2}$ be the expression levels, respectively,$k = 1, \cdots, N$. It is straightforward to calculate the expression variances and their covariance, denoted by $Var(X_1)$, $Var(X_2)$ and $Cov(X_1, X_2)$, respectively. Meanwhile, the last equation of Eq.(1.4) shows that, theoretically, the expression covariance between species (V_{12}) decays exponentially with the component $(\beta_1 + \beta_2)$, providing a foundation to define the expression distance between two species by

$$D_{12} = (\beta_1 + \beta_2)\, t = 2\beta t \tag{1.5}$$

where $\beta = \frac{(\beta_1+\beta_2)}{2}$ is the rate of expression evolution (averaged over two lineages). However, the difficulty to estimate D_{12} is two-folds

- the ancient variance (ρ^2) at root O is usually unknown.
- equations in Eq.(4) are valid only when the optimal expression level (μ) is constant among genes. We address these issues as follows.

3.1.4 Estimation of expression distance under the stationary OU model

We invoke the stationary OU process (sOU) to estimate the unknown ancestral variance ρ^2. It has two assumptions:

1. the evolutionary span (τ) between the tissue ancestor (Z) and the species ancestor (O) is so large that the variance at root O approaches to $\rho^2 = 1/W$;
2. after the speciation, the strength of expression conservation (W) remain constant between species. Under the constant-μ assumption, from Eqs.(1.3) and (1.4) we then have $E[x_1] = E[x_2] = \mu$, $V_{11} = V_{22} = \frac{1}{W}$, as well as the last equation of Eq.(1.4) by

$$V_{12} = \frac{e^{-2\beta t}}{W} \tag{1.6}$$

After calculating the Pearson coefficient of expression correlation, $r_{12} = \frac{Cov(X_1,X_2)}{\sqrt{Var(X_1)Var(X_2)}}$, we obtain a simple formula for estimating the expression distance, that is,

$$D_{12} = -\ln(r_{12}) = -\ln(1 - P_{12}) \tag{1.7}$$

When r_{12} is close to 1 in the case of closely-related species, $D_{12} \approx 1 - r_{12}$, i.e., the Pearson distance $P_{12} = 1 - r_{12}$.

3.1.5 Expression distance estimation when expression optima vary among genes

While expression distance estimation based on Eq. (1.7) is intuitively simple, we have realized that the implied constant-μ assumption is unrealistic. Because the optimal expression level (μ) actually varies considerably among different genes, neglecting this variation may lead to an underestimation. Here we develop a new method to correct this bias.

Suppose that μ varies among genes according to a normal distribution with mean zero and variance V_μ. Under the stationary OU model, we can extend Eq.(1.4) as follows:

the expression variances are given by $V_{11} = V_{22} = 1/W + V_\mu$ and the covariance by

$$V_{12} = \frac{e^{-2\beta t}}{W} + V_\mu \tag{1.8}$$

It follows that the Pearson coefficient of correlation $r_{12} = \frac{V_{12}}{\sqrt{[V11,\ V22]}}$ can be written as

$$r_{12} = \frac{\left(e^{-2\beta t} + WV_\mu\right)}{(1 + W_\mu)} = \pi + (1 - \pi)^{e^{-2\beta t}} \tag{1.9}$$

where $\pi = \frac{WV_\mu}{(1+WV_\mu)}$, which measures the relative $\pi = 0$ means among genes. While $\pi = 0$ means a constant-μ among genes, $\pi = 1$ means indicates a very strong μ variation among genes. $r_{12} \to \pi$ as $t \to \infty$. From Eq.(9) we obtain a general formula to estimate the express distance $D_{12} = 2\beta t$ as follows

$$D_{12} = -\ln\frac{(r_{12} - \pi)}{(1 - \pi)} = -\ln\left[1 - \frac{P_{12}}{1 - \pi}\right] \tag{1.10}$$

where $P_{12} = 1 - r_{12}$ is the Pearson distance. Apparently, Eq.(1.10) is reduced to Eq.(1.7) when $\pi = 0$.

3.1.6 Estimation of π

When Eq.(1.10) is applied in the evolutionary analysis of RNA-seq data, we have to know the parameter π, which can be estimated when RNA-seq data of the same tissue from $n >= 3$ species are available.

Let $x_k = (x_{1k}, \cdots x_{nk})$ be the across-species expression profile of the k-th orthologous gene; $k = 1, \cdots N$.

We first calculate the mean expression level of each k-th orthologous gene, denoted by $x_{.k}$, and then calculate the variance, $Var\,(x_.)$, of $x_{.k}$ over all N genes. It appears that $Var\,(x_.)$ is an asymptotically unbiased estimate of V_μ. Next, the strength of expression conservation (W) can be estimated by a simple method in the following part.

It follows that π can be estimated by

$$\pi = \frac{Var\,(x_.)\,W}{1 + WVar(x_.)} \tag{1.11}$$

3.2 Case Study: Expression Tree Building

In here, we will give an example to build a character tree from expression data (expression phylogeny) applying the above-mentioned methods.

TreeExp can be loaded in usual way:

```
library(TreeExp)
```

We first load the datasets created from six tissues' expression data of nine tetrapod species

```
data("tetraExp")
```

3.2.1 Distance matrix

First, we generate an expression distance matrix of these nine tetrapod species:

```
dismat_pea <- expdist(tetraExp, taxa = "all",
                  subtaxa = "Brain",
                  method = "pea")
as.dist(dismat_pea)
```

```
##                   Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## Chimpanzee_Brain   0.03928696
## Bonobo_Brain       0.05396840       0.02661862
## Gorilla_Brain      0.04542102       0.04346246   0.05571070
## Orangutan_Brain    0.07440973       0.06649678   0.07417594    0.06650759
## Macaque_Brain      0.07097455       0.07025971   0.07569084    0.07142852
## Mouse_Brain        0.14753703       0.14734046   0.14569873    0.15774612
## Opossum_Brain      0.21459736       0.21023241   0.20623679    0.22219040
## Platypus_Brain     0.25207183       0.24840593   0.24225669    0.26264602
## Chicken_Brain      0.27146218       0.27435186   0.26609731    0.27169109
##                   Orangutan_Brain Macaque_Brain Mouse_Brain Opossum_Brain
## Chimpanzee_Brain
## Bonobo_Brain
## Gorilla_Brain
## Orangutan_Brain
## Macaque_Brain         0.07718357
## Mouse_Brain           0.15908149    0.12541882
## Opossum_Brain         0.22389819    0.19697269   0.17580943
## Platypus_Brain        0.26310593    0.23960340   0.22327614    0.20305065
## Chicken_Brain         0.26892835    0.24476930   0.23916935    0.28310733
##                   Platypus_Brain
## Chimpanzee_Brain
## Bonobo_Brain
## Gorilla_Brain
## Orangutan_Brain
## Macaque_Brain
## Mouse_Brain
## Opossum_Brain
## Platypus_Brain
## Chicken_Brain        0.28657956
```

`expdist` function can calculate expression distance directly through the *taxaExo* object. It will extract the expression values and *log*2-transform the values according to the **"taxa"** and **"subtaxa"** specified in the options.

The default method *"pea"* is to calculate pair-wise distances by Pearson distance, which equals 1-Pearson's coefficient of expression level.

Besides *"pea"* , there are numbers of alternative methods to calculate pair-wise expression distances, like*"sou", "sou_v","pea", "spe","euc", "cos", "jsd", "tani", "jac"*.

Of particular, **"sou"** means the general sOU distance method and **"sou_v"** means the special sOU model when expression optima vary among genes that we introduced above.

For comparison, we apply **"pea"**,**"sou"**and **"sou_v"** methods to calculate the paired expression distances between species.

Distance matrix applying **"sou"** method:

```
dismat_sou <- expdist(tetraExp, taxa = "all",
                  subtaxa = "Brain",
                  method = "sou")
as.dist(dismat_sou)
```

```
##                  Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## Chimpanzee_Brain  0.04007952
## Bonobo_Brain      0.05547931       0.02697931
## Gorilla_Brain     0.04648489       0.04443524   0.05732270
## Orangutan_Brain   0.07732362       0.06881087   0.07707107    0.06882245
## Macaque_Brain     0.07361915       0.07284999   0.07870868    0.07410792
## Mouse_Brain       0.15962550       0.15939495   0.15747137    0.17167379
## Opossum_Brain     0.24155877       0.23601657   0.23097009    0.25127352
## Platypus_Brain    0.29044833       0.28555890   0.27741060    0.30468720
## Chicken_Brain     0.31671573       0.32069003   0.30937883    0.31702999
##                  Orangutan_Brain Macaque_Brain Mouse_Brain Opossum_Brain
## Chimpanzee_Brain
## Bonobo_Brain
## Gorilla_Brain
## Orangutan_Brain
## Macaque_Brain         0.08032494
## Mouse_Brain           0.17326052    0.13401016
## Opossum_Brain         0.25347156    0.21936656  0.19335350
## Platypus_Brain        0.30531112    0.27391514  0.25267038    0.22696416
## Chicken_Brain         0.31324380    0.28073202  0.27334448    0.33282915
##                  Platypus_Brain
## Chimpanzee_Brain
## Bonobo_Brain
## Gorilla_Brain
## Orangutan_Brain
## Macaque_Brain
## Mouse_Brain
## Opossum_Brain
## Platypus_Brain
## Chicken_Brain        0.33768435
```

Distance matrix applying **"sou_v"** method:

```
dismat_sou_v <- expdist(tetraExp, taxa = "all",
                 subtaxa = "Brain",
                 method = "sou_v")
dismat_sou_v$pi
```

```
## [1] 0.4321824
```

```
as.dist(dismat_sou_v$distance)
```

```
##                  Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## Chimpanzee_Brain  0.07169946
## Bonobo_Brain      0.09987040       0.04801322
## Gorilla_Brain     0.08337321       0.07963103   0.10326685
## Orangutan_Brain   0.14046409       0.12455399   0.13999038    0.12457556
## Macaque_Brain     0.13352606       0.13208832   0.14306392    0.13444018
## Mouse_Brain       0.30087772       0.30041013   0.29651329    0.32546876
## Opossum_Brain     0.47470846       0.46242662   0.45131469    0.49643952
## Platypus_Brain    0.58686287       0.57531948   0.55625069    0.62092606
## Chicken_Brain     0.65024075       0.66003933   0.63229985    0.65101349
##                  Orangutan_Brain Macaque_Brain Mouse_Brain Opossum_Brain
## Chimpanzee_Brain
## Bonobo_Brain
## Gorilla_Brain
## Orangutan_Brain
## Macaque_Brain         0.14610174
## Mouse_Brain           0.32873050    0.24958856
## Opossum_Brain         0.50139288    0.42601631  0.37051757
## Platypus_Brain        0.62243425    0.54813379  0.49958580    0.44254159
## Chicken_Brain         0.64172713    0.56399841  0.54681221    0.69032820
##                  Platypus_Brain
## Chimpanzee_Brain
## Bonobo_Brain
## Gorilla_Brain
## Orangutan_Brain
## Macaque_Brain
## Mouse_Brain
## Opossum_Brain
## Platypus_Brain
## Chicken_Brain        0.70259882
```

Note that method **"sou_v"** will return an object of *list* which contains two element. The firt one is the estimated value of parameter π. The second one is the distance matrix using **"sou_v"** method.

```
dismat_sou_v$pi
```

```
## [1] 0.4321824
```

Comparing the distance values calculted by the three methods, we can approximately draw a conclusion that while the optimal expression level(μ) actually varies considerably among different genes, neglecting this variation may lead to an underesitmation of expression distance.

Also, if you already have a data frame with normalized expression values, there are internal functions available for creating expression distance matrix.

Function `exptabTE` is helpful to generate an expression level table from the *taxaExp* object and *log*2-transform the values. Also, **"taxa"**" and **"subtaxa"**" can be set by users. For instance,

```
expression_table <- exptabTE(tetraExp, taxa = "all",
                             subtaxa = "Brain")

dismat <- dist.pea(expression_table)
dismat <- dist.sou(expression_table)
dismat <- dist.sou_v(expression_table)
```

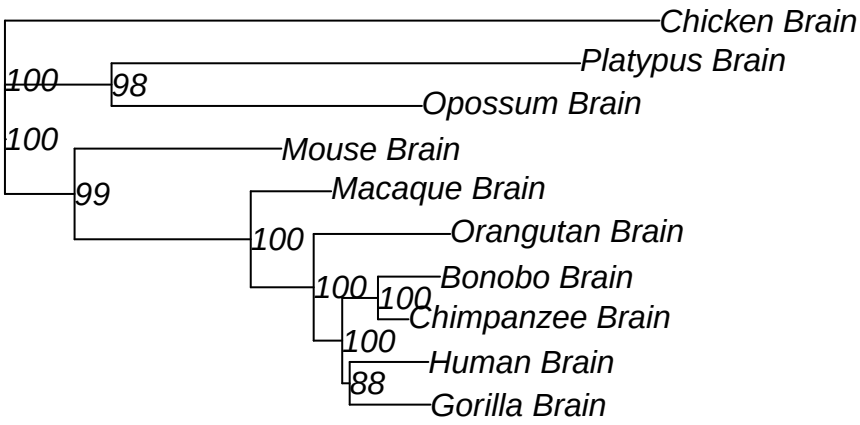
If you have your own expression data frame in the format as the "expression_table" here, it will do fine:

```
dismat <- dist.pea(your_own_dataframe)
colnames(dismat) <- colnames(your_own_dataframe)
rownames(dismat) <- colnames(dismat)
```

3.2.2 Expression character tree

After the expression distance matrix is created, we can construct character tree by Neighbor-Joining method, and bootstrap values based on re-sampling orthologous genes with replacements can also be generated by `boot.exphy` function:

```
library('ape')
 ##### build NJ tree
tr <- NJ(dismat_sou)
 ### estract expression table of Brain tissue from 'tetraExp'
expression_table <-exptabTE(objects = tetraExp,taxa = 'all',
                            subtaxa = 'Brain')
### root the tree using 'Chicken_Brain' as outgroup
tr <- root(tr, "Chicken_Brain", resolve.root = TRUE)
###### generating bootstrap values of the tree
bs<-boot.exphy(phy=tr, x = expression_table, method = 'sou',
               outgroup = 'Chicken_Brain', B = 100)
### assign bootstrap values to the "node.label" of the tree
tr$node.label = bs
### plot the tree
plot(tr, show.node.label = TRUE)
```



Note that the goal of `root` step is to give an better interpretation on the phylogentic relationships of trascriptome data.

Essentially, the tree build by NJ method itself is an unrooted tree. Though `root` function adds a 'root' node to the tree, the branch length between root node and its neighbor internal node is still 0. In other words, `root` step may be only for tree presentation, and it is not needed when a phylogentic tree should be used as input parameter in the following analysis (such like *Ancestral Transcriptome Inference*).

By now, an expression character tree is successfully constructed. The tree shows expression patterns' similarities in selected genes of designated species. The expression tree is largely in accordance with species tree with minor discrepancy.

Phenomenon of evolutionary history dominates the evolutionary expression pattern can be described as phylogenetic signals. One way to interpret highly consistent expression character tree is that expression levels of transcriptome, representing the regulatory changes, accumulated over time. Though not as concrete as sequence data, expression levels generated from transcriptome data across species show strong phylogenetic signals.

Chapter 4

Relative Rate Test for Transcriptome Evolution

4.1 Theory

Given a set (N) of orthologous genes of two species denoted by A and B, respectively, the goal of relative rate test is to investigate whether the rate of transcriptome evolution, on average, differs significantly between two lineages of species A and B. To this end, we need the third species (species C) as outgroup (**Figure 1B**).

Let D_{AB}, D_{AC} and D_{BC} be the pairwise expression distances, respectively, which can be estimated by the methods we formulated above. Let a, b, c be the expression branch lengths of corresponding three lineages A, B and C, respectively. Assuming that these expression distances are additive, we have

$$\begin{aligned} D_{AB} &= a + b \\ D_{AC} &= a + c \\ D_{BC} &= b + c \end{aligned} \tag{2.1}$$

respectively. Since lineages A and B have the same evolutionary time (t), one may write $a = \beta_A t$ and $b = \beta_B t$, where β_A and β_B are the rates of transcriptome evolution in lineages A and B, respectively.

The relative rate test considers the following statistic

$$G_{AB} = D_{AC} - D_{BC} = a - b = (\beta_A - \beta_B)\, t \tag{2.2}$$

Hence, the null hypothesis $G_{AB} = 0$ or $D_{AC} = D_{BC}$ means an equal rate ($\beta_A = \beta_B$) of expression divergence between two lineages. Rejection of this null indicates a rapid expression evolution in lineage A ($\beta_A > \beta_B$ if $G_{AB} > 0$) or in lineage B ($\beta_A > \beta_B$ if $G_{AB} < 0$).

Here, we apply a Z-score test to examine the signigicance

$$Z \;=\; \frac{\triangle_{AB}}{\sqrt{Var\left(\triangle_{AB}\right)}} \tag{2.3}$$

where $\triangle_{AB}$ equals to G_{AB}.

4.2 Statistical Procedure of the Relative Rate Test

4.2.1 Sampling variance of the expression distance

In practice, calculating the sampling variance for the estimated coefficient (r) of correlation is usually carried out using the Fisher transformation, the inverse hyperbolic (artanh) of r, that is, $F(r) = 0.5 \ln\left(\frac{1+r}{1-r}\right)$. It follows that $F(r)$ approximately follows a normal distribution with $F(\rho)$ and the variance of $\frac{1}{N-3}$, where ρ is the true value of the coefficient of correlation and N is the sample size. With the delta method, the inverse Fisher transformation brings the sampling variance back to the correlation scale, resulting in $Var(r) = \frac{\left(1-r^2\right)^2}{N-3}$. Consider the general expression distance given by Eq.(1.11). By the delta method, the large-sampling variance of D_{12} is approximately given by

$$Var(D_{12}) = \frac{Var(r_{12})}{(r_{12} - \pi)^2} \tag{2.4}$$

4.2.2 Calculation of $Var(\triangle_{AB})$

Since $\triangle_{AB} = D_{AC} - D_{BC}$, we have $Var(\triangle_{AB}) = Var(D_{AC}) + Var(D_{BC}) - 2Cov(D_{AC}, D_{BC})$. Moreover, we notice that the sampling covariance equals to $Cov(D_{AC}, D_{BC}) = Var(c)$, as branch-$c$ is the one shared by those two distances. We develop a simple method to calculate $Var(c)$.

1. First, the branch length can be estimated by $c = \frac{(D_{AC}+D_{BC}-D_{AB})}{2}$.
2. we calculate a new variable $r_c = \pi + (1-\pi)e^{-c}$. After viewing r_c as the estimated coefficient of correlation between species C and the ancestral node O, by the Fisher transformation we obtain $Var(c) = \frac{(1-r_c^2)^2}{(N-3)}$. While two sampling variances $Var(D_{AC})$ and $Var(D_{BC})$ can be obtained directly from Eq.(2.4), together we have

$$Var(\triangle_{AB}) = \frac{Var(r_{AC})}{(r_{AC} - \pi)^2} + \frac{Var(r_{BC})}{(r_{BC} - \pi)^2} - \frac{2Var(r_c)}{(r_c - \pi)^2} \tag{2.5}$$

4.3 Case Study: Testing Fast-evolving Genes in Human

In general, relative rate test is to investigate whether the expression of a given gene set in species A is fastly evolved than that in species B. Typically, the gene set chosen to perform the test is set by users and is only a small part of all one-to-one orthologous genes (several hundreds). Moreover, the gene set usually have some specific biological characteristics, such like gene set from a GO term or a co-expression module.

In the case study, we choose gene set with row numbers between 200 to 800 to exemplify how relative rate test works.

At first, we shall extract the gene expression values of brain tissue from the *tetraExp* object.

Parameter `rowindex` defines which genes are selected to perfome the test. Usually, `rowindex` is a vector of numbers corresponded to indices of selecting rows or a vector of logical values (TRUE or FALSE) indicating whether to select the corresponding row or not.

```
#### load the tetraExp data firstly
data(tetraExp)
### extract the gene expression values of the selected genes from 'tetraExp' object.
exp_table <-exptabTE(tetraExp, taxa = 'all', subtaxa = 'Brain',rowindex = 200:800)
```

After obtaining the expression table, performing relative rate test is straightforward by function `RelaRate.test`. It will extract the gene expression value of species A (human), species B (chimpanzee) and outgroup C (macaque) from the *exp_table* and will return a list of three elements. The first one is the Z-score value of the relative rate test, the second is the value of parameter `alternative`, and the last one is the p-value under the specified alternative hypothesis.

```
ztest <- RelaRate.test(expTable = exp_table, x = 'human', y = 'chimpanzee',
                       outgroup = 'macaque', alternative = 'greater')
ztest
```

```
## $Z_score
## [1] 0.7349926
##
## $alternative
## [1] "greater"
##
## $p.value
## [1] 0.231172
```

So, using macaque as an outgroup, the expression of geneset with row numbers between 200 and 800 in human is not likely to evolved significantly faster than that in chimpanzee (p-value >0.05).

Chapter 5

Estimating the Strength of Expression Conservation

5.1 Theory

5.1.1 Stabilizing selection model in transcriptome evolution across species

The Ornstein-Uhlenbeck (OU) model, which claims that expression changes are constrained by the stabilizing selection, is biologically more realistic than a simple Brownian motion (BM) model(Rohlfs et al., 2014, Lemos et al. (2005)Gu and Su (2007)Bedford and Hartl (2009)Brawand et al. (2011)).

Intuitively speaking, the OU model includes two opposite processes:

- random mutations push the expression level (x) of a gene away from the optimum (μ), a process may suffer from a fitness reduction;
- the return process to the expression optimum (μ) driven by a positive selection; the strength of this 'elastic' return increases proportionally with w, the coefficient of stabilizing selection. Given the initial expression value x_0, the OU model predicts that $x(t)$, the expression level after t evolutionary time units, follows a normal distribution with the mean $E[x \mid x_0]$ and variance $V[x \mid x_0]$

$$\begin{aligned} E[x \mid x_0] &= \mu\left(1 - e^{-\beta t}\right) + z_0 e^{-\beta t} \\ V[x \mid x_0] &= \frac{1 - e^{-2\beta t}}{W} \end{aligned} \tag{3.1}$$

respectively, where β is the rate of expression evolution and W is the strength of expression conservation; symbolically one may write an OU process by $OU(x \mid x_0;\ \theta)$, where $\theta = (\mu,\ \beta t,\ W)$ is the parameter vector.

5.1.2 Stationary OU model under a species phylogeny

We shall develop a statistical method to estimate the strength (W) of expression conservation of a gene from high throughput RNA-seq data of multiple species. **Figure 2** illustrates the evolutionary scenario used in the evolutionary transcriptome analysis.

- The first component is the evolutionary lineage from the origin of the tissue (node Z) to the root (node O) of the species phylogeny, with τ evolutionary time units.
- The second component is the conventional species phylogeny with n species. That is, given the initial expression value z_0 at node Z, the OU process of x_0 in the lineage from Z to O is given by $OU\left(x_0 \mid z_0;\ \theta\right)$, where the parameter vector $\theta = (\mu,\ \beta\tau,\ W)$. The joint density of expressions $\boldsymbol{x} = (x_1, ...x_n)$ conditional of the expression level (x_0) at root O, denoted by $P\left(\boldsymbol{x} \mid x_0\right)$, can be derived under the OU model. It follows that the joint expression density of $\boldsymbol{x} = (x_1, ...x_n)$ conditional of z_0 is given by

$$P\left(\boldsymbol{x} \mid z_0\right) = \int_{-\infty}^{\infty} OU\left(x_0 \mid z_0;\ \tau\right) P\left(\boldsymbol{x} \mid x_0\right) dx_0 \tag{3.2}$$

(Hansen and Martins, 1996) showed that either $P\left(\boldsymbol{x} \mid x_0\right)$ or $P\left(\boldsymbol{x} \mid z_0\right)$ is multivariate normally distributed. Consider $P\left(\boldsymbol{x} \mid z_0\right) \sim N\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{V}\right)$ at first, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{V}$ is the variance-covariance matrix. Because current transcriptome data contain little information about the evolution from node Z to node O, calculations of $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are usually difficult because both depend on z_0 and τ.

We propose a stationary OU model (sOU) that helpful to avoid these problems in practice, which postulates that, at the genome-wide level, the biological function of a tissue-specific transcriptome is conservative during the course of species evolution. Specifically, sOU involves two assumptions.

- Origin of the tissue (node Z in **Figure 2**) was so ancient that the evolutionary time between nodes Z and O can be approximated by $\tau \to \infty$. Consequently, the expression mean and variance at root O approach to μ and $\rho^2 = 1/W$, respectively.
- The optimal expression level (μ) and the strength of expression conservation (W) remain virtually constant along the species phylogeny, that is, the expression mean and variances at all internal and external nodes are equal to μ and $1/W$, respectively.

Under the stationary OU model, $P\left(\boldsymbol{x} \mid z_0\right)$ can be simplified as follows:

- $P\left(\boldsymbol{x} \mid z_0\right)$ is independent of z_0;
- the mean vector $\boldsymbol{\mu}$ is uniform, i.e., $\mu_1 = ...\mu_n$;
- the variance-covariance matrix is simply given by $\boldsymbol{V} = \boldsymbol{R}/W$, where $\boldsymbol{R}$ is the coefficient of correlation matrix. Our intent is to estimate the strength of expression conservation of a gene, characterized by a single parameter W. In this sense, the joint density of $\boldsymbol{x}$ can be symbolically written by $P\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{R}, W\right)$. Together, we have

$$P\left(\boldsymbol{x} \mid x_0\right) = P\left(\boldsymbol{x} \mid z_0\right) \sim P\left(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{R},\ W\right) = N\left(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{R}/W\right) \tag{3.3}$$

5.1.3 Variation of W among genes

The sOU model assumes that the strength of expression conservation (W) of a gene remains a constant in species evolution but differs among genes. Substantial evidence has supported this argument(Bedford and Hartl, 2009, Brawand et al. (2011)Cui et al. (2007)Park and Lehner (2013)Tirosh et al. (2006)Warnefors and Eyre-Walker (2012)Zou et al. (2011)). Further, we model W as a random variable that varies among genes according to a gamma distribution, that is,

$$\phi\left(W; \alpha, \overline{W}\right) = \frac{(\alpha/\overline{W})^{\alpha}}{\Gamma\left(\alpha\right)} W^{\alpha-1} e^{-\alpha W/\overline{W}} \tag{3.4}$$

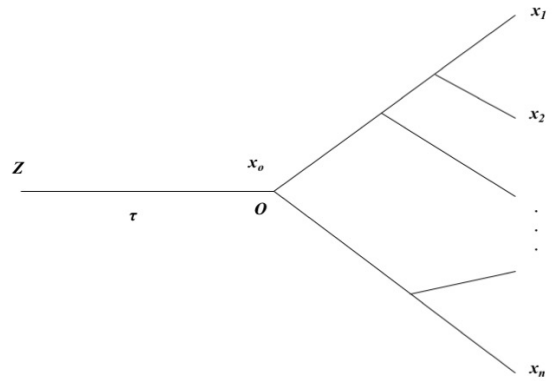


Figure 5.1: Figure2. The evolutionary scenario for comparative transcriptome analysis

where $\overline{W}$ is the mean and α is the shape parameter; a small values of α means a high degree of W-variation, and $\alpha = \infty$ means a constant W among genes. After rewriting the joint normal density $P(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{R},\ W)$ by $P(\boldsymbol{x} \mid W;\ \boldsymbol{\mu},\ \boldsymbol{R},)$ to indicate W is a random variable, we have

$$P(\boldsymbol{x} \mid W;\ \boldsymbol{\mu},\ \boldsymbol{R},) = N(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{R}/W) \tag{3.5}$$

It follows that the marginal density of $\boldsymbol{x}$ is given by

$$\begin{aligned} P\left(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{R},\alpha,\overline{W}\right) &= \int_0^\infty P(\boldsymbol{x}|W;\ \boldsymbol{\mu};\boldsymbol{R})\,\phi\left(W;\ \alpha,\overline{W}\right)dW \\ &= A\left(\frac{\overline{W}}{\alpha}\right)^{n/2}\left(\frac{\Gamma(n/2+\alpha)}{\Gamma(\alpha)}\right)\left(\frac{\alpha}{\alpha+Q(\boldsymbol{x})\,\overline{W}}\right)^{n/2+\alpha} \end{aligned} \tag{3.6}$$

where $Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^T\,\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is a quadratic function of $\boldsymbol{x}$, and $A = \pi^{-\frac{n}{2}}\,|\boldsymbol{R}|^{-\frac{1}{2}}$ is a normalization constant.

5.1.4 An empirical Bayesian framework for gene-specific W esitimation

5.1.4.1 Posterior mean of W as gene-specific predictor

We adopt an empirical Bayesian procedure to predict the strength of expression conservation for single gene. By the Bayes rule, the posterior density of W conditional of the expression profile ($\boldsymbol{x}$) of a gene is given by

$$P\left(W \mid \boldsymbol{x};\boldsymbol{\mu},\boldsymbol{R},\alpha,\overline{W}\right) = \frac{\phi\left(W;\alpha,\overline{W}\right)P(\boldsymbol{x}|W;\boldsymbol{\mu},\boldsymbol{R})}{P\left(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{R},\alpha,\overline{W}\right)} \tag{3.7}$$

After some mathematical calculations, one can show that the analytical form of the posterior density of W is given by

$$P\left(W \mid \boldsymbol{x};\boldsymbol{\mu},\boldsymbol{R},\alpha,\overline{W}\right) = A\left(\frac{\overline{W}}{\alpha}\right)^{n/2}\left(\frac{\Gamma(n/2+\alpha)}{\Gamma(\alpha)}\right)\left(\frac{\alpha}{\alpha+Q(\boldsymbol{x})\,\overline{W}}\right)^{n/2+\alpha} \tag{3.8}$$

Hence, $P\left(W \mid \boldsymbol{x};\boldsymbol{\mu},\boldsymbol{R},\alpha,\overline{W}\right)$ follows a gamma distribution, with the mean and variance given by

$$\begin{aligned} E[W|\boldsymbol{x}] &= \left[\frac{\alpha+\frac{n}{2}}{\alpha+Q(x)\,\overline{W}}\right]\overline{W} \\ Var[W|\boldsymbol{x}] &= \left[\frac{\alpha+\frac{n}{2}}{\left(\alpha+Q(x)\,\overline{W}\right)^2}\right](\overline{W})^2 \end{aligned} \tag{3.9}$$

respectively. Of particular, the posterior mean, $E[W|\boldsymbol{x}]$ can be used as the predictor for the strength of expression conservation of a gene with obseved expression profile $\boldsymbol{x}$.

5.1.4.2 Relative strength of expression conservation

It has been realized that the strenth of expression conservation (W_k) of gene k highly depends on the normalization method used for RNA-seq raw reads count. Hence, it is difficult to compare between two sets of estimates when they used different normalization methods. To aleviate this problem, it is more convenient to use the ratio $U_k = W_k/\overline{W}$, the relative strength of expression conservation. Suppose we have N orthologous genes under study, and the expression profile of the k-th gene is denoted by $\boldsymbol{x}_k$, $k = 1, N$. Let $W_k = E[W|\boldsymbol{x}]$ be the posterior predictor for the strength of expression conservaton of gene k. Notice that the expectation of the posterior mean prediction ($E[W|\boldsymbol{x}]$) with respect to the the maginal density $P\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{R}, \alpha, \overline{W}\right)$ is equal to the mean of the strength of expression conservation ($\overline{W}$), that is,

$$\int E[W \mid \boldsymbol{x}]\, P\left(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{R},\ \alpha,\ \overline{W}\right) d\boldsymbol{x} = \overline{W} \tag{3.10}$$

Eq.(3.10) implies that the averge of U_k, the relative strength of expression conservation over all genes is roughly to be one, that is,

$$\frac{\sum_{k=1}^{N} U_k}{N} \approx 1 \tag{3.11}$$

5.2 Statistical Procedure

Suppose that we have RNA-seq datasets of a particular tissue from n species, and the expression profile of each k-th gene denoted by $\boldsymbol{x_k} = (x_{1k}, ...x_{nk})$, $k = 1..., N$. We developed a practically feasible procedure to estimate W of each gene, which actually deals with the quadratic function of $\boldsymbol{x_k}$, or $Q(\boldsymbol{x_k})$. The procedure is briefly described below.

1. Calculate gene-k specific mean (μ_k) by a simple average over orthologous genes.
2. Calculate the matrix of correlation coefficients ($\boldsymbol{R}$) from comparative RNA-seq data, which is applied to each of gene.
3. Calculate the quadratic function of each gene k, $Q(x_k)$, by

$$\hat{Q}(x_k) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(x_{ik} - \mu_k)(x_{jk} - \mu_k) \tag{3.12}$$

   where c_{ij} is the ij-th element of matrix $\boldsymbol{C} = \boldsymbol{R}^{-1}$, x_{ik} or x_{jk} is the expression value of gene k in species i or j, respectively.
4. Treating $Q(\boldsymbol{x_k})$ as the observation of gene k and rewriting Eq.(6), symbolically, by $P\left(Q(\boldsymbol{x_k}); \alpha, \overline{W}\right)$, build up an approximate likelihood function

$$\ell\left(X \mid \alpha, \overline{W}\right) = \prod_{k=1}^{N} P\left(\hat{Q}(\boldsymbol{x_k}); \alpha, \overline{W}\right)$$

   and obtain the maximum likelihood estimates (MLE) of α and $\overline{W}$; the standard likelihood ratio test is applied to test the null hypothesis of no W-variation among genes, i.e., $\alpha = \infty$.
5. Calculate W_k, the empirical Bayesian estimate of the strength of expression conservation of gene k, from Eq.(3.9) after replacing α and $\overline{W}$ by their estimates.

5.3 Case Study: Emprical Bayesian Esitimates of Expression Conservation of genes

We use the expression values of 5635 1:1 orthologous genes in brain of nine mammalian species to estimate the parameters of the selection pressure gamma distribution in brain. Then we estimate the gene-specific selection pressure based on Bayes' theorem.

TreeExp can be loaded the package in the usual way:

```
library('TreeExp')
```

Let us first load the tetrapod expression dataset:

```
data('tetraExp')
```

5.3.1 Inversed correlation matrix

And then, based on the constructed *taxaExp* object, we are going to create an inverse correlation matrix between mammalian species from the *taxaExp* object:

```
species.group <- c("Human", "Chimpanzee", "Bonobo", "Gorilla",
"Macaque", "Mouse", "Opossum", "Platypus")
### all mammalian species

inv.corr.mat <- corrMatInv(tetraExp, taxa = species.group, subtaxa = "Brain")
inv.corr.mat
```

```
##                  Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## Human_Brain       15.7867153      -7.05168768   -0.5467655   -5.21892601
## Chimpanzee_Brain  -7.0516877      26.74465629  -13.4479799   -4.73381009
## Bonobo_Brain      -0.5467655     -13.44797986   18.3508826   -1.88760497
## Gorilla_Brain     -5.2189260      -4.73381009   -1.8876050   14.49822033
## Macaque_Brain     -1.9749665      -1.07677087   -1.3431113   -2.65851421
## Mouse_Brain       -0.6505556      -0.21643982   -0.3682007    0.27736989
## Opossum_Brain     -0.1345412      -0.31082672   -0.1962514   -0.02468375
## Platypus_Brain    -0.1113688      -0.01113911   -0.4318870   -0.01712743
##                  Macaque_Brain Mouse_Brain Opossum_Brain Platypus_Brain
## Human_Brain         -1.9749665  -0.6505556   -0.13454125    -0.11136876
## Chimpanzee_Brain    -1.0767709  -0.2164398   -0.31082672    -0.01113911
## Bonobo_Brain        -1.3431113  -0.3682007   -0.19625137    -0.43188695
## Gorilla_Brain       -2.6585142   0.2773699   -0.02468375    -0.01712743
## Macaque_Brain        9.9086664  -2.0792122   -0.50552772    -0.26332694
## Mouse_Brain         -2.0792122   5.1274351   -1.27896910    -0.61687114
## Opossum_Brain       -0.5055277  -1.2789691    3.90792798    -1.19664014
## Platypus_Brain      -0.2633269  -0.6168711   -1.19664014     3.02761767
```

5.3.2 Estimation of gamma parameters

Then we need to extract the expression values of orthologous genes from the *taxaExp* object using `exptabTE` function.

```
brain.exptable <- exptabTE(tetraExp, taxa = species.group, subtaxa = "Brain" ,logrithm = TRUE)
head(brain.exptable)
```

```
##                 Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## ENSG00000198824   4.6111724         5.029011     5.151778     4.1093606
## ENSG00000118402   5.3652726         5.597233     5.360715     4.9064096
## ENSG00000166167   6.6794801         6.687621     7.174127     6.4891254
## ENSG00000144724   4.8374387         4.356144     5.460087     4.2280490
## ENSG00000183508   0.9855004         1.327687     2.788686     0.9030383
## ENSG00000008086   4.9321557         4.516015     5.825277     4.7213727
##                 Macaque_Brain Mouse_Brain Opossum_Brain Platypus_Brain
## ENSG00000198824      5.861707    6.700994      7.218781       6.536675
## ENSG00000118402      5.909533    7.396091      7.014913       8.321252
## ENSG00000166167      7.304967    8.257011      7.593279       6.903038
## ENSG00000144724      5.852249    6.657211      7.164605       6.239551
## ENSG00000183508      1.395063    3.382667      3.240314       2.989139
## ENSG00000008086      7.000113    6.889352      7.595966       7.036943
```

With the inverse correlation matrix and expression values of brain tissue in 9 mammals, we are now able to estimate the parameters of the gamma distribution:

```
gamma.paras <- estParaGamma(exptable =brain.exptable, corrmatinv =inv.corr.mat)
## print the elements of gamma.paras
gamma.paras
```

```
## $alpha
## [1] 2.937128
##
## $W_average
## [1] 0.2610562
##
## $speNum
## [1] 8
##
## $geneNum
## [1] 5636
```

The $\bar{W}$ is the average of the selection pressure levels in the tissue brain. And the shape parameter α here can reflect the internal variances of selection pressure. The more close α is to 2, the more distinctive selection pressures on genes. And if the α is close to infinite, it means there are no difference among selection pressures on genes.

5.3.3 Bayesian estimation of gene-specific selection pressure

After parameters of the gamma distribution are estimated, we are able to estimate posterior selection pressures as well as their *se* with given 'RPKM' values across species:

```
brain.Q <- estParaQ(brain.exptable, corrmatinv = inv.corr.mat)
# with prior expression values and inversed correlation matrix

brain.post<- estParaWBayesian(brain.Q, gamma.paras)
```

```
brain.W <- brain.post$w # posterior selection pressures
brain.CI <- brain.post$ci95 # posterior expression 95% confidence interval
```

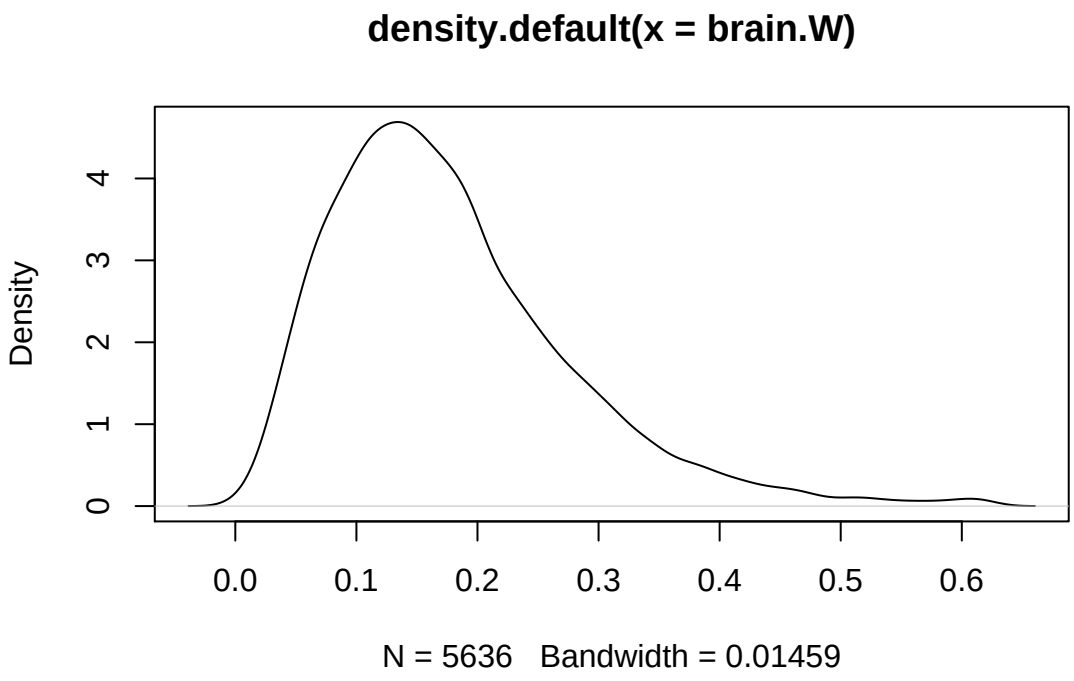
After esitimating the Bayesian selection pressres, W, we can chek a few gene with hightest selection pressure.

```
names(brain.W) <- rownames(brain.exptable)
head(sort(brain.W, decreasing = TRUE)) #check a few genes with highest seletion pressure
```

```
## ENSG00000137270 ENSG00000102243 ENSG00000139515 ENSG00000146378
##        0.616582        0.616582        0.616582        0.616582
## ENSG00000151379 ENSG00000111049
##        0.616582        0.616582
```

and draw the dentisty plot of W among genes.

```
plot(density(brain.W))
```



Chapter 6

Ancestral Transcriptome Inference

6.1 Theory

The empirical Bayesian method for ancestral expression inference along a phylogeny is illustrated in **Figure 3A**. While the theoretical foundation has been well formulated by the work of (Gu, 2004) based on the Brownian motion (BM) model, some updates and refinements are necessary before the Ornstein-Uhlenbeck (OU) model that is biologically more realistic,e.g.,(Bedford and Hartl, 2009).

6.1.1 From Brownian Motion (BM) model to Ornstein-Uhlenbeck (OU) model

Let X be the expression level of a gene. Brownian motion is the simplest model to describe changes in gene expression between orthologous genes, in which the degree of stochastic change away from the current state is independent of both state and time. It is well-known that the probability density function of a Brownian motion is given by

$$p\left(x \mid x_0,\ \sigma^{2t}\right) \tag{4.1}$$

where x_0 is the state of the process at time 0. In the evolutionary context, σ^2 describes the rate of expression divergence only driven by the mutational effects, which corresponds to the case of selective neutrality.

As the basic model, the notion of optimal expression claims that stabilizing selection, which maintains the optima under the background of random mutations, dominates the transcriptome evolution (Hansen and Martins, 1996). For instance, under-expression of metabolic enzymes may slow the metabolic flux, while over-expression may expose the cell to additional toxic misfolded proteins. Following the most common practice, the stabilizing selection on the expression of a gene (x) satisfies a Gaussian-like fitness,

$$f\left(x\right) = e^{-\frac{\omega(x-\mu)^2}{2}} \tag{4.2}$$

where μ is the optimal value, ω is the coefficient of stabilizing selection; a large ω means a strong selection pressure, and vice versa. Treating $f\left(x\right)$ as the cost function, one can mathematically show that the evolution of X follows an Ornstein-Uhlenbeck (OU) stochastic process. That is, given the initial expression value \$x_\$0, the OU model predicts that $x\left(t\right)$, the values of X after t evolutionary time units, follows a normal distribution with the following mean $E\left[x|x_0\right]$ and variance $V\left(x|x_0\right)$

$$\begin{aligned} E\left[x|x_0\right] &= \mu\left(1-e^{-\beta t}\right) + x_0 e^{-\beta t} \\ V\left(x|x_0\right) &= 1 - \frac{e^{-2\beta t}}{W} \end{aligned} \tag{4.3}$$

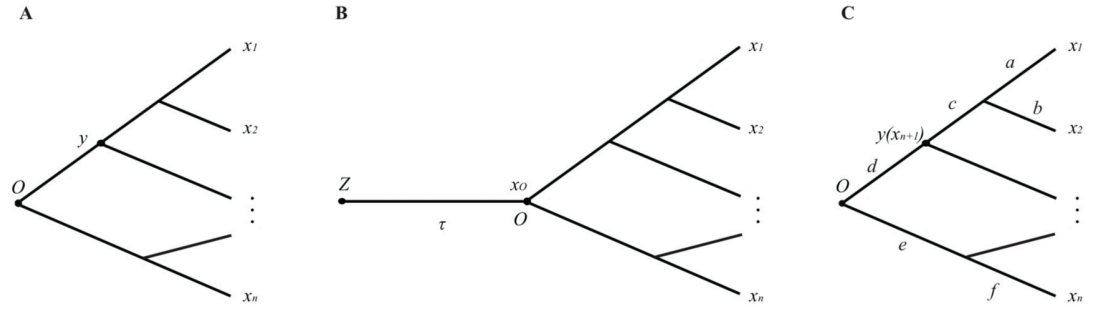


Figure 6.1: Figure3. The schematic of ancestral expression inference along a phylogeny

respectively, where the rate of expression evolution $\beta = W\sigma 2$, and $W = 4Ne\omega$; Ne is the effective population size(Gu, 2004, Hansen and Martins (1996)). Hence, an OU model can be concisely represented by $OU\,(x \mid x_0;\ t,\ \beta,\ W)$. Intuitively speaking, an OU process can be thought of as adding an elastic spring to a Brownian motion. As random mutations push the gene expression farther away from this fixed optimum, the strength of elastic return increases proportionally.

6.1.2 The OU model under a phylogeny

Given the expression level (x_0) at the root O of a known phylogeny with n species, the joint density of expression levels of a gene, $\boldsymbol{x} = (x_1, \cdots, x_n)$, is denoted by $P(\boldsymbol{x})$. It has been shown (Hansen and Martins, 1996) that it follows a multi-variate normal distribution, that is, $P(\boldsymbol{x}) = N(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{V})$, or explicitly

$$P(\boldsymbol{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |\boldsymbol{V}|^{\frac{1}{2}}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{V}^{\{-1\}} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right\} \tag{4.4}$$

where the mean vector ($\boldsymbol{\mu}$) and the variance-covariance matrix ($\boldsymbol{V}$) depends on the phylogenetic structure, branch lengths, rates of expression evolution and the initial expression level (x_0) at root O.

6.1.3 Ancestral gene expression inference: an empirical Bayesian approach

Our approach provides an empirical Bayesian procedure to infer the expression states of ancestral nodes along a given phylogeny. Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be the observed expression profile and y be that at any ancestral node of interest (**Figure 3A**). According to the Bayes rule, the posterior density $P(y \mid x_1 ..., x_n)$ is computed as follows

$$P(y \mid x_1, \cdots,\ x_n) = \frac{P(x_1,\ \cdots,\ x_n,\ y)}{P(x_1,\ \cdots,\ x_n)} \tag{4.5}$$

This claims that $P(x_1,\ \cdots,\ x_n)$ is an n-variate normal density, denoted by $N(x_1,\ \cdots,\ x_n, \boldsymbol{\mu}, \boldsymbol{V})$. Let $M = n+1$ and regard the ancestral level y as an additional variable $xn+1$. Under the BM model, Gu(Gu, 2004) showed that $P(x_1,\ \cdots,\ x_n,\ y)$ is an $(n+1)$-variate normal density, denoted by $N(x_1,\ \cdots,\ x_n, \boldsymbol{\mu}, \boldsymbol{V_M})$. The extended variance-covariance matrix V_M has the following structure:

$$\boldsymbol{V_M} = \begin{bmatrix} V & \boldsymbol{H} \\ \boldsymbol{H}' & V_{n+1,n+1} \end{bmatrix} \tag{4.6}$$

where $\boldsymbol{H} = (H_1,\ \cdots,\ H_n)^T$. That is, for $i,\ j = 1,\ \cdots n$, the ij-th element of $\boldsymbol{V_M}$ is equal to that of $\boldsymbol{V}$. For any $i = 1,\ \cdots n+1$, the element $V_{i,n+1} = V_{n+1,i} = H_i$, respectively. It is straightforward to show that

these results holds under the OU model, except for the details of $\boldsymbol{V_M}$ elements. It follows that the posterior density $P(y \mid x_1, \cdots, x_n)$ is a normal density. Let $\boldsymbol{C} = \boldsymbol{V_M}^{-1}$; c_{ij} is the ij-th element of $\boldsymbol{C}$. After some algebras we obtain

$$P(y \mid x_1, \cdots, x_n) = \frac{1}{\sqrt{2\pi\sigma_{y|x}}} \exp\left(-\frac{1}{2\sigma^2_{y|x}} \left[y - \mu + \sum_{i=1}^{N} \frac{c_{i,n+1}}{c_{n+1,n+1}} (x_i - \mu) \right]^2 \right) \tag{4.7}$$

where $\sigma^2_{y|x} = \frac{1}{c_{n+1,n+1}}$ is the (posterior) variance of y.

6.1.4 The problem of ancestral transcriptome inference

Under the stabilizing selection model that the transcriptome evolution follows an Ornstein-Uhlenbeck (OU) stochastic process, the problem is to infer the expression level (y) of an ancestral of interest along a given phylogeny (**Figure 3A**). Through the empirical Bayesian procedure, we have shown that the posterior density $P(y \mid x_1, \cdots, x_n)$is given by Eq.(4.7), and the the posterior mean of y conditional of $x_1, \cdots, x_n$ can be given by

$$y \mid x = E[y \mid x_1, \cdots, x_n] = b_0 + \sum_{i=1}^{n} b_i x_i \tag{4.8}$$

where $b_i = -\frac{c_{i,n+1}}{c_{n+1,\ n+1}}$and $b_0 = \mu(1 + b_1 + \cdots + b_n)$. Hence, it can be used as an empirical Bayesian predictor for the ancestral state of gene expression. In short, the ancestral expression inference is a simple linear combination of the current expressions; their (linear) coefficients are determined by the phylogeny and the model parameters.

From the view of practice, it appears that the problem of ancestral expression inference is virtually to determine those linear coefficients that are associated with each ancestral node of interest ($b_1, \cdots, b_n$). However, when the general OU model is applied, some computational difficulties may arise, mainly due to the model complexity that causes the problem of over-parametrization. To this end, we introduce the stationary OU model (sOU), under which we develop a feasible procedure to calculate the coefficients b_i, $i = 0, \cdots, n$.

6.1.5 Stationary Ornstein-Uhlenbeck (sOU) model

Let z_0 be the expression level at the birth time (node Z) of a particular tissue (**Figure 3C**). Since then, the evolutionary change of x_0 along the lineage follows an OU model given by $OU(x_0 \mid z_0; \tau, \beta_0, W)$. The stationary OU model (sOU) invokes two assumptions.

1. The timing of tissue origin was much more ancient than the root of species phylogeny under study such that node O can be approximated by the stationary condition of $OU(x_0 \mid z_0; \tau, \beta_0, W)$ as $\tau \to \infty$. In other words, the mean and the variance of x_0 at the root of phylogeny is simply given by μ and $\rho^2 = 1/W$, respectively.
2. Since then, the optimal level (μ) and the strength of stabilizing selection (W) remain constant during the course of evolution along the species phylogeny. Consequently, the expression variances in all internal and external nodes are the same, which equal to $1/W$. Under the stationary OU model, it can be shown that the variance-covariance matrix $\boldsymbol{V}$ is root-independent(Hansen and Martins, 1996).

6.2 A Fast Algorithm under the Stationary OU Model

The algorithm to calculate the coefficients $b_0,\ b_1,\ \cdots,\ b_n$ under the stationary OU model can be briefed as follows.

- As the expression variance is expected to be the same among all species, the expression variance (V_0) is the simple average of expression variances among species.
- The coefficient of correlation between the i-th and j-th external nodes, denoted by R_{ij}, is calculated by the standard approach.
- Let $\boldsymbol{R}$ be the matrix of coefficients of correlation. It is straightforward to calculate the variance-covariance matrix of $P(x_1, \cdots, x_n)$ by $\boldsymbol{V} = V_0\boldsymbol{R}$.

- The difficulty in calculating the variance-covariance matrix of $P(x_1, \cdots, x_n, y)$, $\boldsymbol{V_M}$, is how to calculate the covariance elements between the i-th external node and the $(n+1)$-th (internal) node, $i = 1, \cdots, n$ (**Figure 3C**). Under the stationary OU model, we show $V_{i,n+1} = V_0 e^{-di}$, where d_i is the expression branch length from the internal node (y) to the i-th external node.

We develop a simple method, to estimate d_i by mapping the expression distance matrix onto the known phylogeny. As $R_{n+1,i} = R_{i,n+1}$ and $R_{n+1,n+1} = 1$, we then obtain $\boldsymbol{V_M}$ and its inverse matrix $\boldsymbol{C}$.

6.3 Case Study: Ancestral Expression Inference of Gene *EMP1*

In here, we will walk through an example of how to perform ancestral expression estimation on primates' expression data of brain tissue. The test dataset inluce brain expression values from six primates speceis which are Human, Chimpanzee, Bonobo, Gorilla, Orangutan and Macaque.

6.3.1 Expression character tree

Note that the presumption of ancestral state inference is that the transcriptome datasets must contain sufficient phylogentic signals. An empirical approach to verifying this assumption is to compare the inferred tree with the species tree.

We first generate brain expression distance matrix of these six primate species:

```
data('tetraExp')
primate_group <-c('human','chimpanzee','gibbon','bonobo','gorilla','macaque')

dismat <- expdist(tetraExp, taxa = primate_group,
                  subtaxa = "brain", method = "sou")
dismat
```

```
##                  Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## Human_Brain       0.00000000       0.04007952   0.05547931    0.04648489
## Chimpanzee_Brain  0.04007952       0.00000000   0.02697931    0.04443524
## Bonobo_Brain      0.05547931       0.02697931   0.00000000    0.05732270
## Gorilla_Brain     0.04648489       0.04443524   0.05732270    0.00000000
## Macaque_Brain     0.07361915       0.07284999   0.07870868    0.07410792
##                  Macaque_Brain
## Human_Brain         0.07361915
## Chimpanzee_Brain    0.07284999
## Bonobo_Brain        0.07870868
## Gorilla_Brain       0.07410792
## Macaque_Brain       0.00000000
```

Alternatively, distance matrix can also be generated via a twostep procedure. First is to exstract the expression data of self-defined primate species and then calculate the pairwise distance of the expression data.

The two ways to construct distance matrix come to the same results.

```
primate_expT <-exptabTE(objects = tetraExp,taxa = primate_group,
                            subtaxa = 'Brain')
dismat <- dist.sou(primate_expT)
dismat
```

```
##                  Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## Human_Brain       0.00000000       0.04007952   0.05547931    0.04648489
## Chimpanzee_Brain  0.04007952       0.00000000   0.02697931    0.04443524
## Bonobo_Brain      0.05547931       0.02697931   0.00000000    0.05732270
## Gorilla_Brain     0.04648489       0.04443524   0.05732270    0.00000000
## Macaque_Brain     0.07361915       0.07284999   0.07870868    0.07410792
##                  Macaque_Brain
## Human_Brain         0.07361915
```

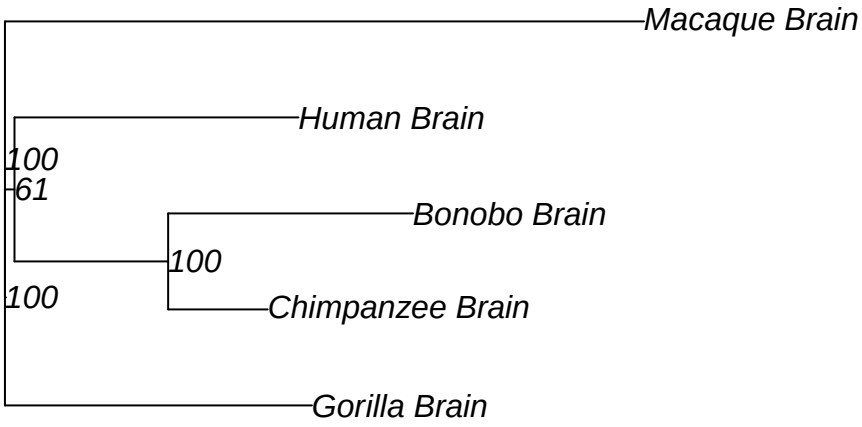
```
## Chimpanzee_Brain    0.07284999
## Bonobo_Brain        0.07870868
## Gorilla_Brain       0.07410792
## Macaque_Brain       0.00000000
```

Next, we build the NJ-tree based on the distance matrix we have just constructed.

```
primate_tree <- NJ(dismat)
```

To compare the built expression tree to the species tree, we can root the tree with respect to the specified outgroup ('Macaque_Brain' here) and plot the tree.

```
primate_tree_root <- root(primate_tree, outgroup = "Macaque_Brain", resolve.root = TRUE)
bs<-boot.exphy(phy=primate_tree_root, x = primate_expT, method = 'sou', outgroup = 'Macaque_Brain',B = 
primate_tree_root$node.label = bs
plot(primate_tree_root, show.node.label = TRUE)
```



Bootstrap values is helpful to confirm the accuracy of the infered tree. From the tree we built, we can approximatly conclude the the inferred trees are virtually consistant with the know species phylogeny with high bootstrap (53-100) supports.

6.3.2 Creating variance co-variance matrix

In this step, we will calculate the variance-covariance matrix of $P(x_1, \cdots, x_n, y)$ by $\boldsymbol{V_M} = V_0 \boldsymbol{R_M}$. Since $V_{i,n+1} = V_0 e^{-di}$ where d_i is the expression branch length from the internal node (y) to the i-th external node and could be estimated by mapping the expression distance matrix onto the known phylogeny, the primate expression tree built in the previous step is necessary here as input data.

Ultimately, we are going to get the inverse matix $\boldsymbol{C}$ of variance-covariance matrix $\boldsymbol{V_M}$ applying `varmatInv` function.

```
var_mat <- varMatInv(objects = tetraExp,phy = primate_tree,
                     taxa = primate_group, subtaxa = "Brain")
var_mat
```

```
##               1             2             3             4             5
## 1  3.761503e+00  3.839017e-14  1.863141e-14  1.718778e-15  2.986023e-16
## 2  4.002279e-14  1.057909e+01  1.801250e-15 -4.149628e-14  2.731195e-15
```

```
## 3  5.181866e-15 -1.487986e-15  4.350370e+00 -8.931328e-15  5.002194e-15
## 4  3.293732e-15 -4.201812e-14 -8.821499e-15  3.487037e+00  8.128807e-15
## 5  8.848801e-16  1.866274e-14  1.180825e-14  2.844059e-15  1.718434e+00
## 6 -2.530769e-13  2.738326e-12  6.736726e-13 -3.404190e+00 -1.634523e+00
## 7 -3.678729e+00 -2.820728e-12 -7.347222e-13 -3.490486e-13  4.510490e-14
## 8 -6.186680e-14 -1.049691e+01 -4.267722e+00  9.005267e-14 -2.995133e-15
##               6             7             8
## 1 -2.133443e-13 -3.678729e+00 -7.212699e-14
## 2  2.707509e-12 -2.782327e-12 -1.049691e+01
## 3  2.558188e-13 -2.720751e-13 -4.267722e+00
## 4 -3.404190e+00  1.114175e-14  6.874319e-14
## 5 -1.634523e+00  5.085846e-14 -4.585997e-14
## 6  1.145600e+02 -1.096000e+02 -3.622996e-12
## 7 -1.096000e+02  1.199982e+02 -6.799842e+00
## 8 -3.886726e-12 -6.799842e+00  2.148421e+01
```

6.3.3 Ancestral expression estimation

Here, we extract the brain expression values of six primates species, locate *EMP1*(ENSEMBL ID: ENSG00000105695) gene and extract its expression value vector:

```
primate_expT <-exptabTE(objects = tetraExp,taxa = primate_group,
                            subtaxa = 'Brain')

EMP1_expression <- primate_expT[which(rownames(primate_expT) == "ENSG00000134531"),]
```

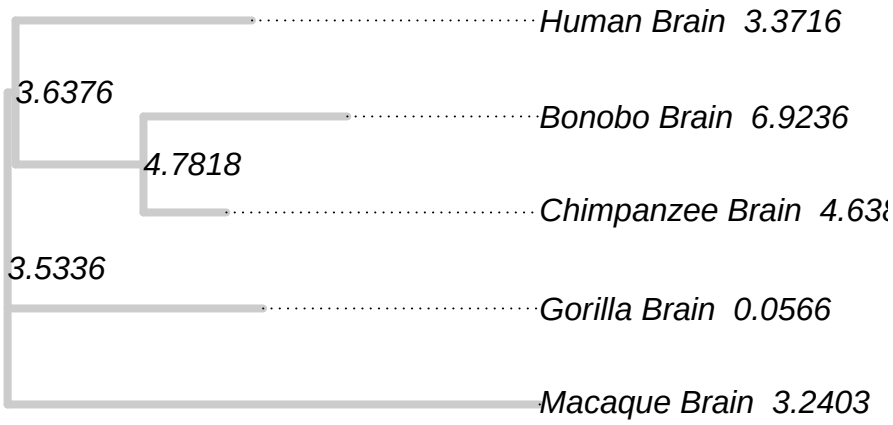
Then, function `aee` will estimate the posterior mean of y conditional of $x_1, \cdots, x_n$ based on Eq.(4.8).

```
EMP1_anc <- aee(x = EMP1_expression, phy = primate_tree, mat = var_mat)
```

Finally, we map these estimations on the primate expression tree to give a direct presentation of these values:

```
primate_tree$node.label <- sprintf("%.4f",EMP1_anc$est)
primate_tree$tip.label <- paste0(primate_tree$tip.label, "  ",
                                 sprintf("%.4f", EMP1_expression))

plot(primate_tree, edge.color = "grey80", edge.width = 4,
     show.node.label = TRUE, align.tip.label = TRUE)
```



Bibliography

Bedford, T. and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences*, 106:1133–1138.

Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Biology*, 2:e9.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478:343.

Butler and King (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164:683.

Cui, Q., Yu, Z., Purisima, E. O., and Wang, E. (2007). MicroRNA regulation and interspecific variation of gene expression. *Trends in Genetics*, 23:372–375.

Gu, X. (2004). Statistical Framework for Phylogenomic Analysis of Gene Family Expression Profiles. *Genetics*, 167:531–542.

Gu, X., Ruan, H., and Yang, J. (2019). Estimating the Strength of Expression Conservation from High Throughput RNA-seq Data. *Bioinformatics*.

Gu, X. and Su, Z. (2007). Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences*, 104:2779–2784.

Hansen, T. F. and Martins, E. P. (1996). Translating Between Microevolutionary Process and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data. *Evolution*, 50:1404.

Lemos, B., Meiklejohn, C. D., Cáceres, M., and Hartl, D. L. (2005). Rate of divergence in gene expression profiles of primates, mice and flies: stabilizing selection and variability among functionnal categories. *Evolution*, 63:126.

Park, S. and Lehner, B. (2013). Epigenetic epistatic interactions constrain the evolution of gene expression. *Molecular Systems Biology*, 9:645.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140.

Rohlfs, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation. *Molecular Biology and Evolution*, 31:201–211.

Ruan, H., Su, Z., and Gu, X. (2016). Treeexp1.0: R package for analyzing expression evolution based on rna-seq data. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 326(7):394–402.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nature Genetics*, 38:ng1819.

Warnefors, M. and Eyre-Walker, A. (2012). A Selection Index for Gene Expression Evolution and Its Application to the Divergence between Humans and Chimpanzees. *PLoS ONE*, 7:e34935.

Xu, C., Li, Q., Efimova, O., He, L., Tatsumoto, S., Stepanova, V., Oishi, T., Udono, T., Yamaguchi, K., Shigenobu, S., Kakita, A., Nawa, H., Khaitovich, P., and Go, Y. (2018). Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome Research*, 28:1097–1110.

Yang, J., Ruan, H., Zou, Y., Su, Z., and Gu, X. (2018). Ancestral transcriptome inference based on rna-seq and chip-seq data. *Methods*.

Zou, Y., Huang, W., Gu, Z., and Gu, X. (2011). Predominant Gain of Promoter TATA Box after Gene Duplication Associated with Stress Responses. *Molecular Biology and Evolution*, 28:2893–2904.