Bayesian estimation of gene-specific conservation levels

Hang Ruan

2017/3/28

It is reasonable to assume that the stabilizing selection pressures on expression levels of different genes within a species should be different. In here, we assume that the selection pressures on genes in a tissue within a species follow a gamma distribution:

$$\phi(W) = \frac{(\alpha/\bar{W})^{\alpha}}{\Gamma(\alpha)} W^{\alpha-1} e^{-\alpha W/\bar{W}}$$

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')
```

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", "Orangutan",
                   "Macaque", "Mouse", "Opossum", "Platypus")
### all mammalian species

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

Estimation of gamma parameters

Then we need to extract the 'RPKM' values of orthologous genes from the *taxaExp* object.

```
brain.exptable <- exptabTE(tetraexp.objects, taxa = species.group, subtaxa = "Brain" )
head(brain.exptable)
```

```
##                 Human_Brain Chimpanzee_Brain Bonobo_Brain Gorilla_Brain
## ENSG00000198824   4.6111274         5.028821     5.151915     4.1096027
## ENSG00000118402   5.3654189         5.597289     5.360793     4.9061705
## ENSG00000166167   6.6794593         6.687562     7.174159     6.4891530
## ENSG00000144724   4.8375065         4.356123     5.459962     4.2283255
## ENSG00000183508   0.9826653         1.329606     2.788397     0.9067318
## ENSG00000008086   4.9322604         4.516231     5.825182     4.7215191
##                 Orangutan_Brain Macaque_Brain Mouse_Brain Opossum_Brain
## ENSG00000198824        4.509800      5.861672    6.701035      7.218736
## ENSG00000118402        5.380957      5.909652    7.396126      7.014964
## ENSG00000166167        6.309653      7.304929    8.257017      7.593249
```

```
## ENSG00000144724        4.065409      5.852348    6.657203      7.164571
## ENSG00000183508        0.834059      1.394680    3.382727      3.241028
## ENSG00000008086        4.197305      7.000130    6.889332      7.595977
##                  Platypus_Brain
## ENSG00000198824       6.536603
## ENSG00000118402       8.321238
## ENSG00000166167       6.903008
## ENSG00000144724       6.239608
## ENSG00000183508       2.988988
## ENSG00000008086       7.036899
```

With the inverse correlation matrix and 'RPKM' values, we are now able to estimate the parameters of the gamma distribution:

```
gamma.paras <- estParaGamma(brain.exptable, inv.corr.mat)
cat(gamma.paras)
```

```
## 3.317864 0.2154153 59.80488 7688.483 2.149649 9 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.

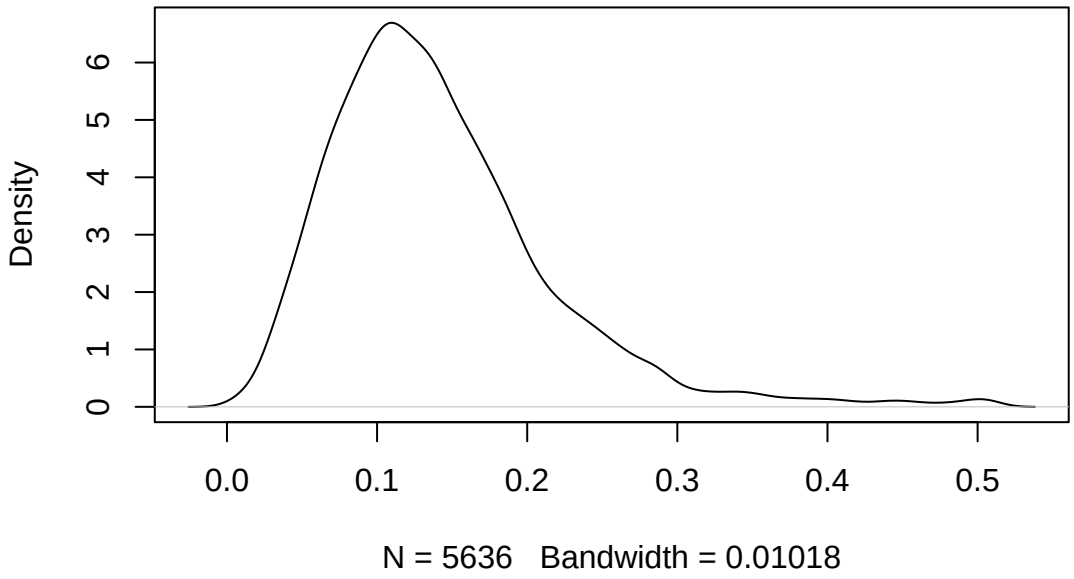
Bayesian estimation of gene-specific selection pressure

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$exp # posterior expression values
brain.CI <- brain.post$ci95 # posterior expression 95% confidence interval

names(brain.W) <- rownames(brain.exptable)

head(sort(brain.W, decreasing = T)) #check a few genes with highest seletion pressure
```

```
## ENSG00000137270 ENSG00000102243 ENSG00000139515 ENSG00000146378
##       0.5075818       0.5075818       0.5075818       0.5075818
## ENSG00000151379 ENSG00000111049
##       0.5075818       0.5075818
```

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


density.default(x = brain.W)
Density
0
1
2
3
4
5
6
0.0
0.1
0.2
0.3
0.4
0.5
N = 5636 Bandwidth = 0.01018