

# Report: Case Studies in Bioinformatics, Module 1

*Cyril Matthey-Doret*

*1 janvier 2017*

## Background

Ontogeny is the development of an organism from egg fertilization to its mature state. It requires expression of different genes at each stage of development, associated with phenotypic changes. There is a lot of discussion on how these sets of genes differ between species. The earliest studies of embryogenesis performed by Karl Ernst Ritter von Baer in the 19th century suggested that the most similar state of development between related species, also known as the phylotypic stage, happens at the earliest stages of embryogenesis. This led to what is called the funnel model. However, more recent work has shown that this convergence happens later, earliest stages being more different between species, giving rise to the hourglass model. It is now possible to test these models by combining gene expression data at different stages of embryogenesis and phylogenetic information for the concerned genes to better study the differences in expressed genes between species. Here, I investigate a study relying on an approach termed phylostratigraphy, which consists in tracing the origin of specific genes through similarity searches. This allows to build a phylogenetic tree where each gene has a rank that estimates their age. These ranks can be used to get a metric of age for the whole transcriptome at each developmental stage, called Transcriptome Age Index (TAI). Results obtained in this study support the hourglass model, focusing on expression datasets from the zebrafish *Danio rerio*, *Drosophila*, the mosquito *Anopheles* and the nematode *Caenorhabditis elegans*. In this report, I use forensic bioinformatics to repeat the analysis from Domazet-Lošo et Tautz, 2010 to have a critical view of each step and a better understanding of their results. Finally, I propose improvements for the analysis, explaining how it would affect the results.

## Methods

I reproduced the analysis from the paper using python 2 with the modules panda, numpy and GEOparse. I used expression data from *Danio rerio* (zebrafish) to reproduce the results shown in figure 1a of the paper. Expression data for zebrafish was obtained from Domazet-Lošo et Tautz (2010) and phylostrata were obtained from Šestak et al, 2013. The first step was to format the expression data from microarray and find the gene ID corresponding to the spots ID in the microarray. This information can be obtained via the GEO platform (GPL) annotation file.

```
## Parsing ./GSE24616.soft.gz:
## - DATABASE : GeoMame
## - SERIES : GSE24616
## - PLATFORM : GPL6457
## - SAMPLE : GSM606866
## - SAMPLE : GSM606867
## - SAMPLE : GSM606868
## - SAMPLE : GSM606869
## - SAMPLE : GSM606870
## - SAMPLE : GSM606871
## - SAMPLE : GSM606872
## - SAMPLE : GSM606873
## - SAMPLE : GSM606874
## - SAMPLE : GSM606875
## - SAMPLE : GSM606876
## - SAMPLE : GSM606877
## - SAMPLE : GSM606878
```

## - SAMPLE : GSM606879  
## - SAMPLE : GSM606880  
## - SAMPLE : GSM606881  
## - SAMPLE : GSM606882  
## - SAMPLE : GSM606883  
## - SAMPLE : GSM606884  
## - SAMPLE : GSM606885  
## - SAMPLE : GSM606886  
## - SAMPLE : GSM606887  
## - SAMPLE : GSM606888  
## - SAMPLE : GSM606889  
## - SAMPLE : GSM606890  
## - SAMPLE : GSM606891  
## - SAMPLE : GSM606892  
## - SAMPLE : GSM606893  
## - SAMPLE : GSM606894  
## - SAMPLE : GSM606895  
## - SAMPLE : GSM606896  
## - SAMPLE : GSM606897  
## - SAMPLE : GSM606898  
## - SAMPLE : GSM606899  
## - SAMPLE : GSM606900  
## - SAMPLE : GSM606901  
## - SAMPLE : GSM606902  
## - SAMPLE : GSM606903  
## - SAMPLE : GSM606904  
## - SAMPLE : GSM606905  
## - SAMPLE : GSM606906  
## - SAMPLE : GSM606907  
## - SAMPLE : GSM606908  
## - SAMPLE : GSM606909  
## - SAMPLE : GSM606910  
## - SAMPLE : GSM606911  
## - SAMPLE : GSM606912  
## - SAMPLE : GSM606913  
## - SAMPLE : GSM606914  
## - SAMPLE : GSM606915  
## - SAMPLE : GSM606916  
## - SAMPLE : GSM606917  
## - SAMPLE : GSM606918  
## - SAMPLE : GSM606919  
## - SAMPLE : GSM606920  
## - SAMPLE : GSM606921  
## - SAMPLE : GSM606922  
## - SAMPLE : GSM606923  
## - SAMPLE : GSM606924  
## - SAMPLE : GSM606925  
## - SAMPLE : GSM606926  
## - SAMPLE : GSM606927  
## - SAMPLE : GSM606928  
## - SAMPLE : GSM606929  
## - SAMPLE : GSM606930  
## - SAMPLE : GSM606931  
## - SAMPLE : GSM606932

## - SAMPLE : GSM606933  
## - SAMPLE : GSM606934  
## - SAMPLE : GSM606935  
## - SAMPLE : GSM606936  
## - SAMPLE : GSM606937  
## - SAMPLE : GSM606938  
## - SAMPLE : GSM606939  
## - SAMPLE : GSM606940  
## - SAMPLE : GSM606941  
## - SAMPLE : GSM606942  
## - SAMPLE : GSM606943  
## - SAMPLE : GSM606944  
## - SAMPLE : GSM606945  
## - SAMPLE : GSM606946  
## - SAMPLE : GSM606947  
## - SAMPLE : GSM606948  
## - SAMPLE : GSM606949  
## - SAMPLE : GSM606950  
## - SAMPLE : GSM606951  
## - SAMPLE : GSM606952  
## - SAMPLE : GSM606953  
## - SAMPLE : GSM606954  
## - SAMPLE : GSM606955  
## - SAMPLE : GSM606956  
## - SAMPLE : GSM606957  
## - SAMPLE : GSM606958  
## - SAMPLE : GSM606959  
## - SAMPLE : GSM606960  
## - SAMPLE : GSM606961  
## - SAMPLE : GSM606962  
## - SAMPLE : GSM606963  
## - SAMPLE : GSM606964  
## - SAMPLE : GSM606965  
## - SAMPLE : GSM606966  
## - SAMPLE : GSM606967  
## - SAMPLE : GSM606968  
## - SAMPLE : GSM606969  
## - SAMPLE : GSM606970  
## - SAMPLE : GSM606971  
## - SAMPLE : GSM606972  
## - SAMPLE : GSM606973  
## - SAMPLE : GSM606974  
## - SAMPLE : GSM606975  
## - SAMPLE : GSM606976  
## - SAMPLE : GSM606977  
## - SAMPLE : GSM606978  
## - SAMPLE : GSM606979  
## - SAMPLE : GSM606980  
## - SAMPLE : GSM606981  
## - SAMPLE : GSM606982  
## - SAMPLE : GSM606983  
## - SAMPLE : GSM606984  
## - SAMPLE : GSM606985  
## - SAMPLE : GSM606986

## - SAMPLE : GSM606987  
## - SAMPLE : GSM606988  
## - SAMPLE : GSM606989  
## - SAMPLE : GSM606990  
## - SAMPLE : GSM606991  
## - SAMPLE : GSM606992  
## - SAMPLE : GSM606993  
## - SAMPLE : GSM606994  
## - SAMPLE : GSM606995  
## - SAMPLE : GSM606996  
## - SAMPLE : GSM606997  
## - SAMPLE : GSM606998  
## - SAMPLE : GSM606999  
## - SAMPLE : GSM607000  
## - SAMPLE : GSM607001  
## - SAMPLE : GSM607002  
## - SAMPLE : GSM607003  
## - SAMPLE : GSM607004  
## - SAMPLE : GSM607005  
## - SAMPLE : GSM607006  
## - SAMPLE : GSM607007  
## - SAMPLE : GSM607008  
## - SAMPLE : GSM607009  
## - SAMPLE : GSM607010  
## - SAMPLE : GSM607011  
## - SAMPLE : GSM607012