Can Ascl1 alone predict phylogeny of organisms from Animalia? Comparing results from Decipher and Clustal Omega

Mrinal Vashisth, M4235

*Report submitted to professor **Mike Ryako**, as a project for **Phylogenetics module, M4235, ITMO, 2019***

Abstract

Ascl1 (achaete-scute family bHLH transcription factor 1) is involved in development of cerebral cortex. In this study 249 orthologues of Ascl1 were mined from NCBI, and subsequently analysed using Decipher and Clustal Omega. Phylogenetic analysis and clustering were done with same parameters. And finally the results were put into a broader perspective. It was found that indeed this gene can be used to plot a robust phylogenetic tree. Clustering results from Decipher turned out to be more accurate in this analysis as compared to Clustal Omega. In both speed and accuracy Decipher outperformed Clustal Omega suggesting effectiveness of Bayesian Dynamic Programming approach.

Introduction

Body size to neocortex proportion in humans is greatest among animals. This also gives humans unique mental capacities otherwise present to a lesser extent in nearest neighbors, the primates. Genes linked within development specific complex networks are termed as 'hubs' (Greenhill et al. 2017). One such gene is Ascl1. The selection of this gene was done using Gene Ontology Annotations for the GO:0021987. The particular result is as follows:

```
MGI:96919      Ascl1   achaete-scute family bHLH transcription factor 1    10
cerebral cortex development  results in the development of cerebral cortex,has the
participant oligodendrocyte          IMP     MGI:1857470    J:132162
```

The evidence is stated as inferred from mutants. As can be seen in this interaction image from Huri database ("Interactome (protein Interactome)" 2015), it is a complex network, and these are only validated and verified connections.

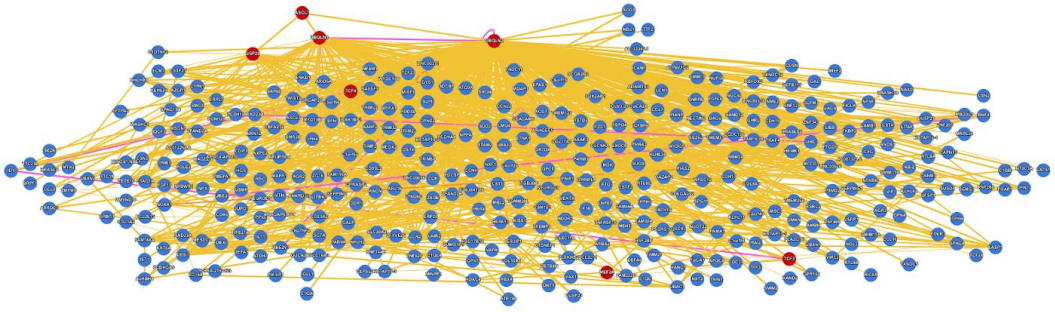


Fig. 1 Huri interactome network for Ascl1 with key interactors USP20, UBQLN2, TCF4, TCF3, MEF2A, UBQLN1. Click here for the full size image.

Multiple sequence alignment (MSA) in R

MSA fits given sequences in a form that reflects shared quality of the underline sequences. Decipher is a dynamic programming method trained to maximise scoring matrices for making structural/ evolutionary alignments. It is also a fast algorithm, the same algorithm took less than 30 seconds to align 249 sequences as compared to 5 minutes taken by Clustal Omega. Even seen in the original benchmarks, and confirmed by the results of this study Decipher is not fairly accurate as compared to similar competitor algorithms (Wright 2015). Clustal Omega uses Guide Trees and HMM profiles to perform the alignment. It has impressive speed benchmarks ((Wright 2015; Sievers and Higgins 2014).

Two MSA libraries were used in this study: `DECIPHER`, and `msa`. The code can be accessed from here.

Steps

Step 1: NCBI search was done using rentrez::search for the following query

```
(ASCL1[All Fields] AND orthologs[All Fields]) AND alive[prop]
```

However with this approach we are getting also transcript variants. Therefore we use regex expression

Step 2: Gene ids were used with rentrez::fetch to get fasta format sequences for 249 results.
Step 3: Decipher and Clustal Omega were ran for the samples.
Step 4: Phylogenetic analyses and clustering were done along with bootstrapping.
Step 5: Validation were based by comparing with another standardized tree.

Results

Base of tree for Decipher consists of Pan and Rattus genus suggesting that these are the most advanced genes in the entire alignment. While the base of Clustal Omega consists of the same species but with finer division. Additionally the Rattus branch is salient, alongside Gallus. But Xenopus is also incorrectly assigned to the same tree. Additionally, different species of

Mus are assigned different clades within the same level.

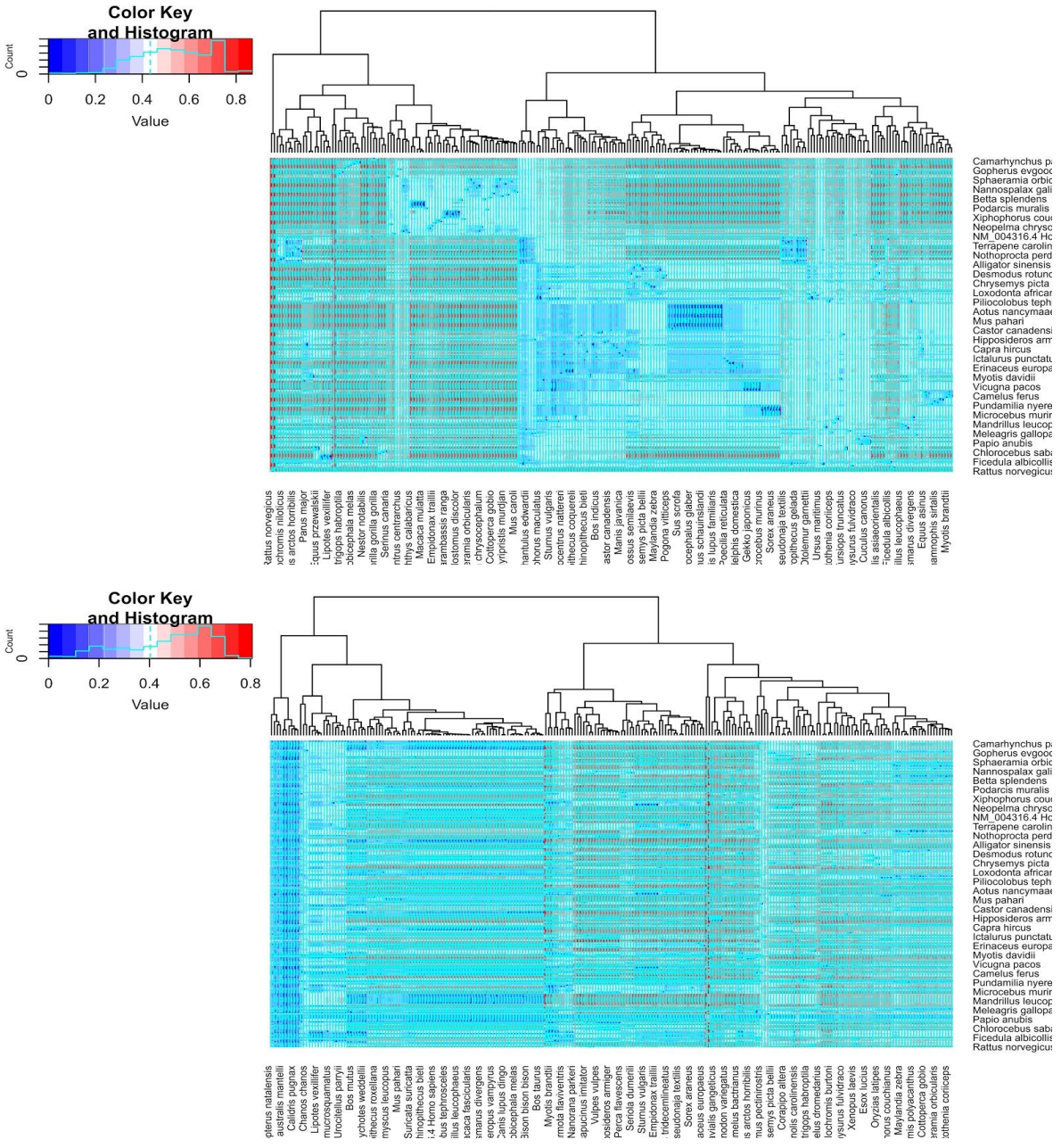


Fig 2. Upper: Heatmap of distance matrix and dendrogram for DECIPHER, Lower: Distance matrix heatmap for Clustal Omega alignment distance matrix. Neighbour joining method was used for distance calculation.

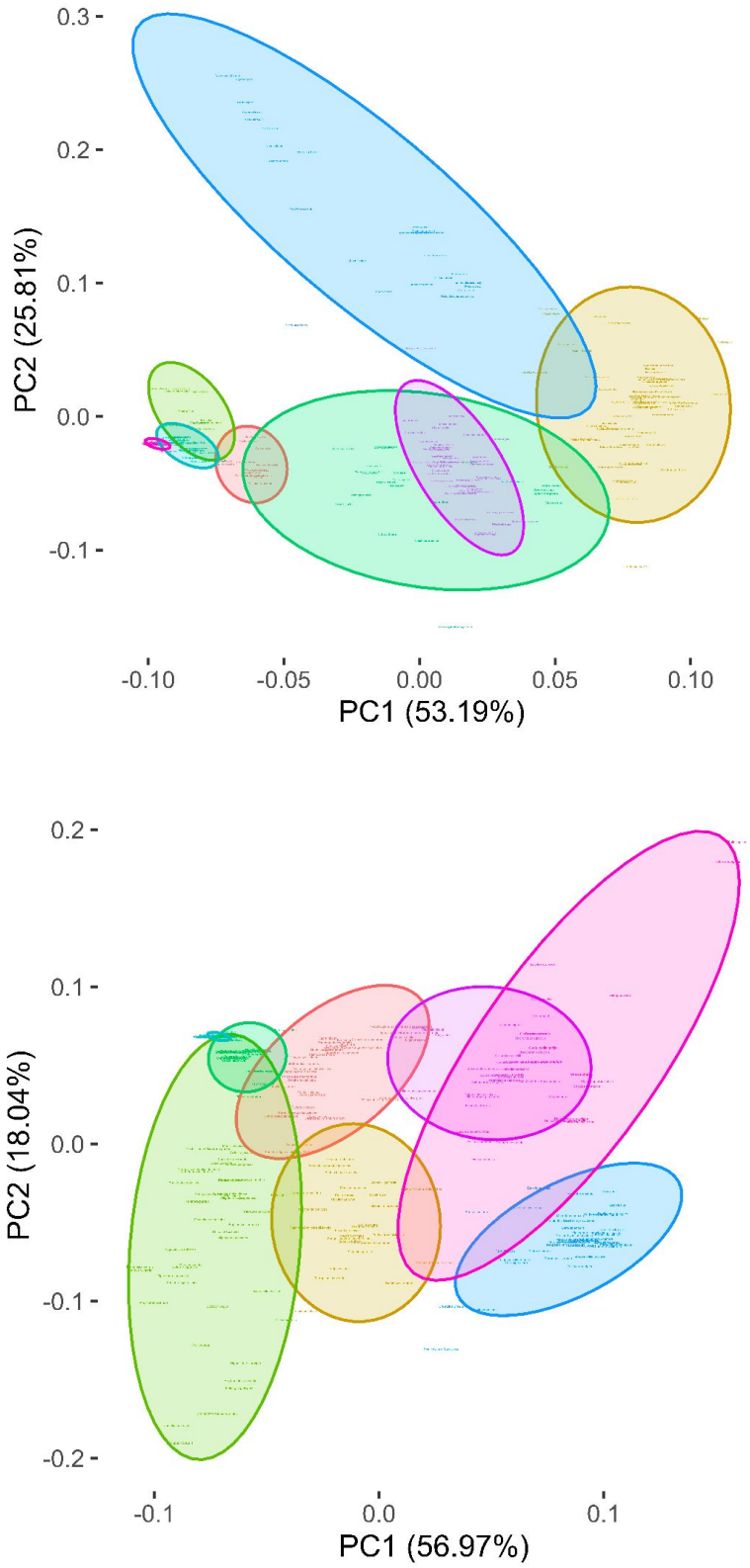


Fig 3. K-means clusters for 8 clusters from alignment results from DECIPHER (above) and Clustal Omega (below).

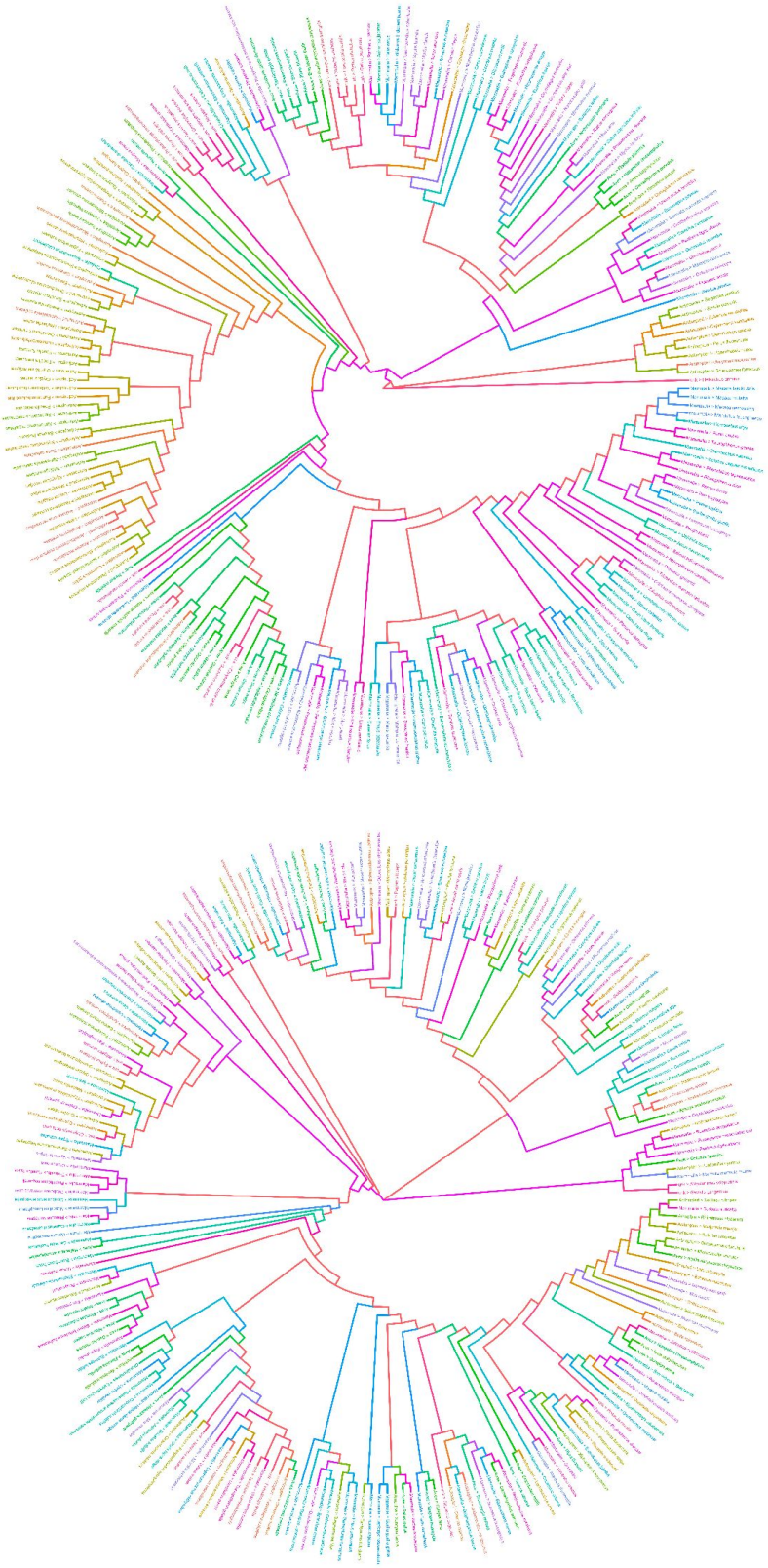
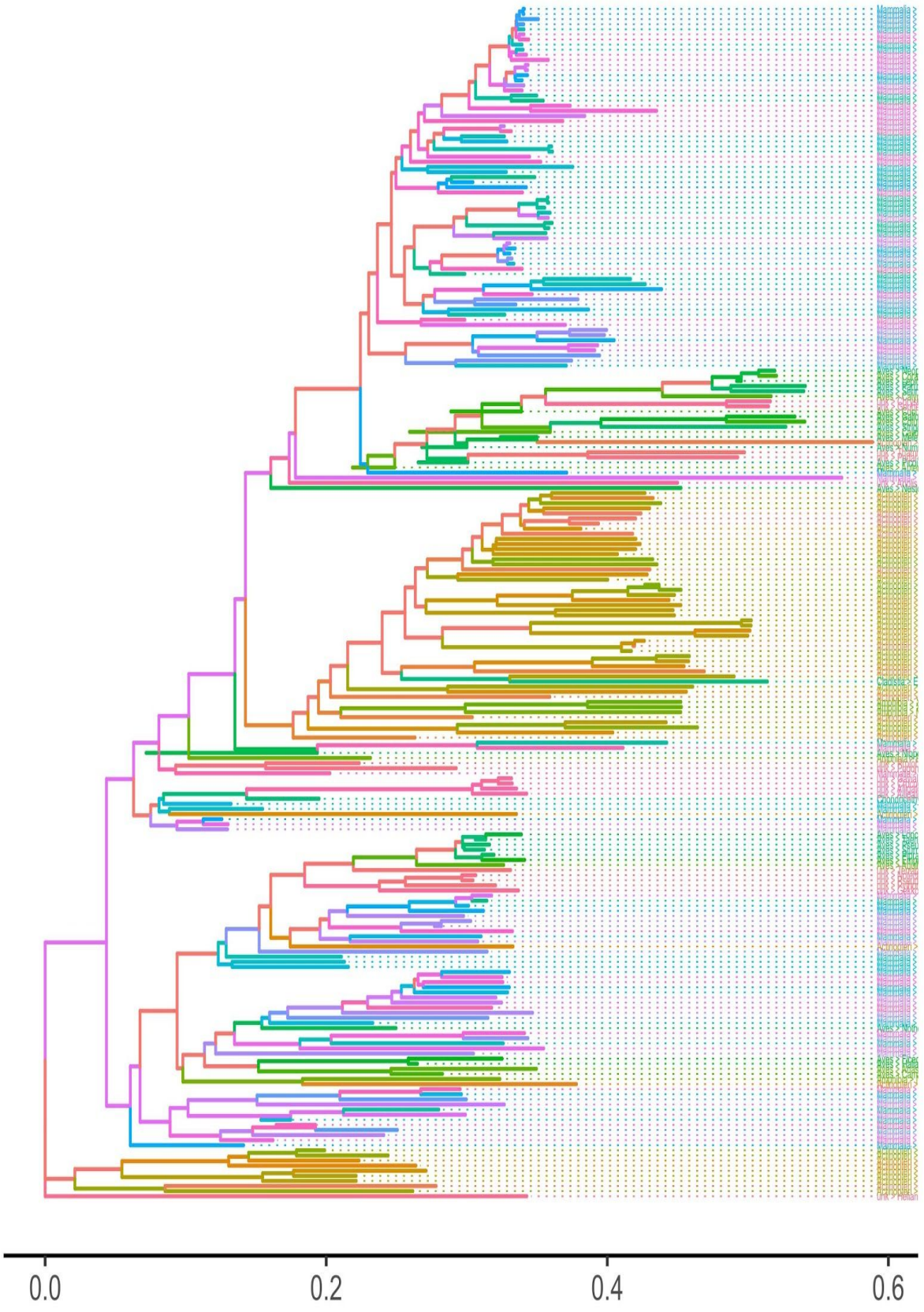


Fig 4. Phylogenetic trees from DECIPHER (above), Clustal Omega (below). Both images put Helianthus as a separate branch. The colouring is done at class level information. Mammalia form a neat clade in DECIPHER, shown in orange colour.


0.0
0.2
0.4
0.6

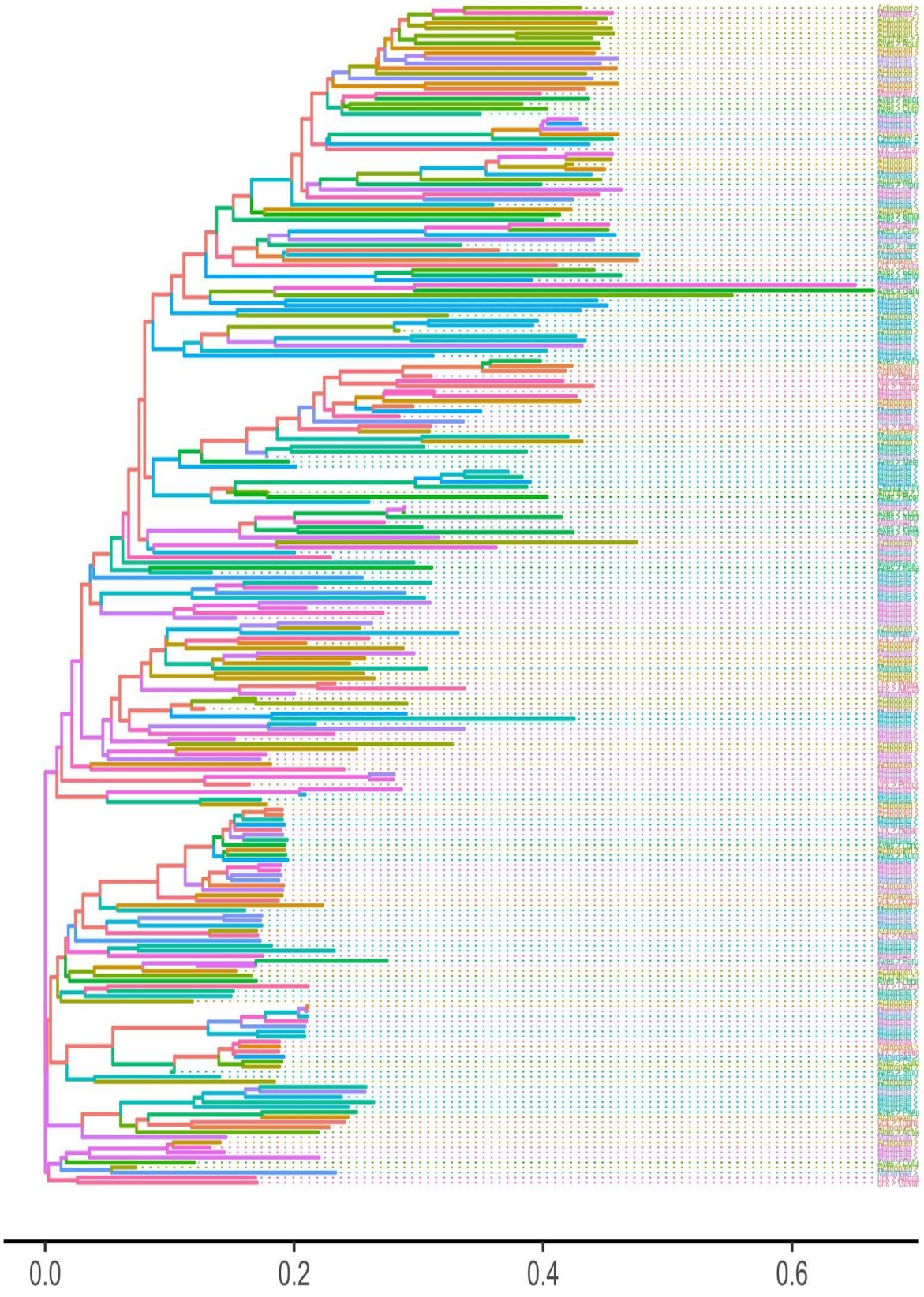


Fig 5. Rooted trees with Helianthus group as the root for DECIPHER (above), and Clustal Omega (below).

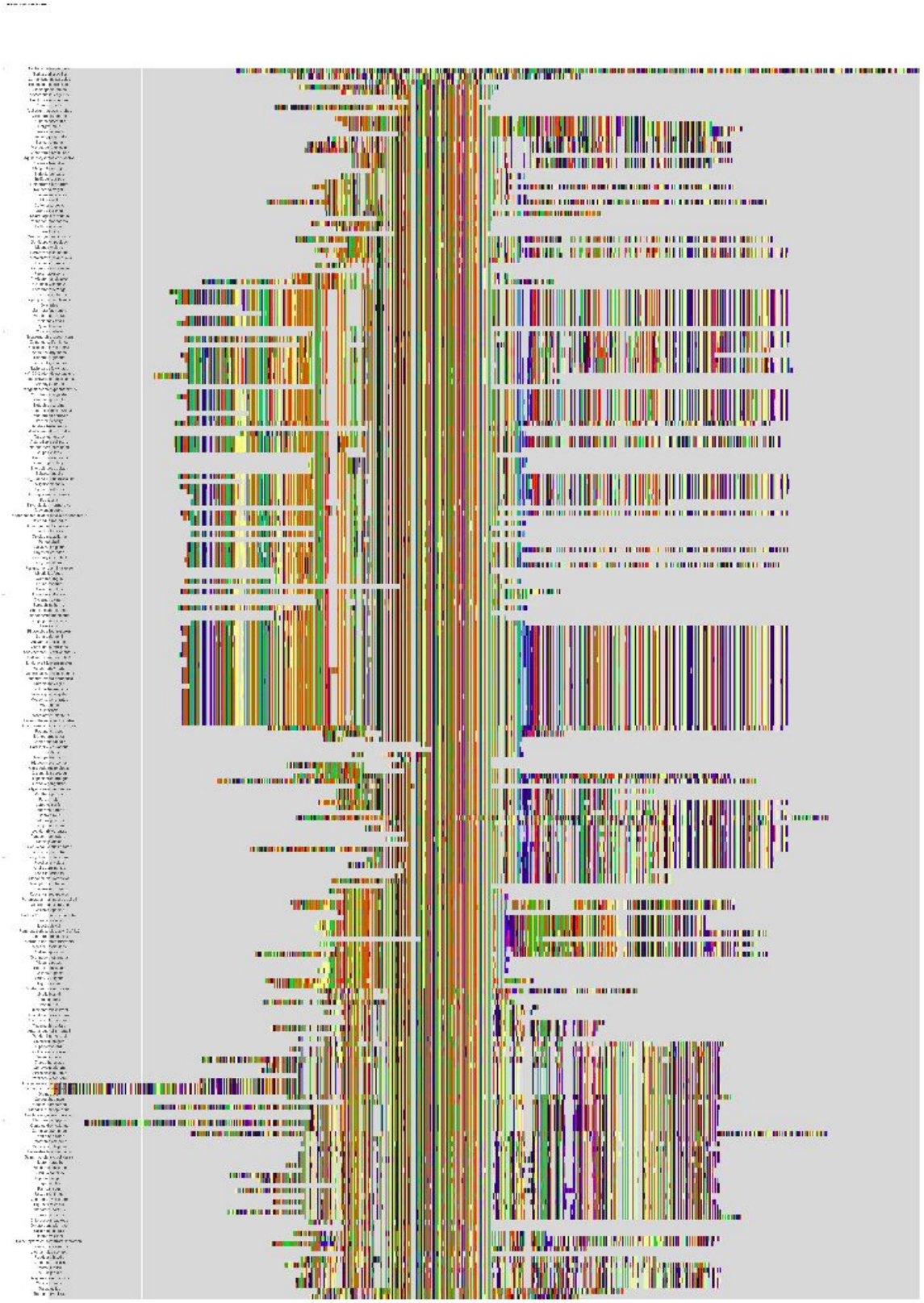


Fig 6. Alignment plots for DECIPHER (above) and Clustal Omega (below).

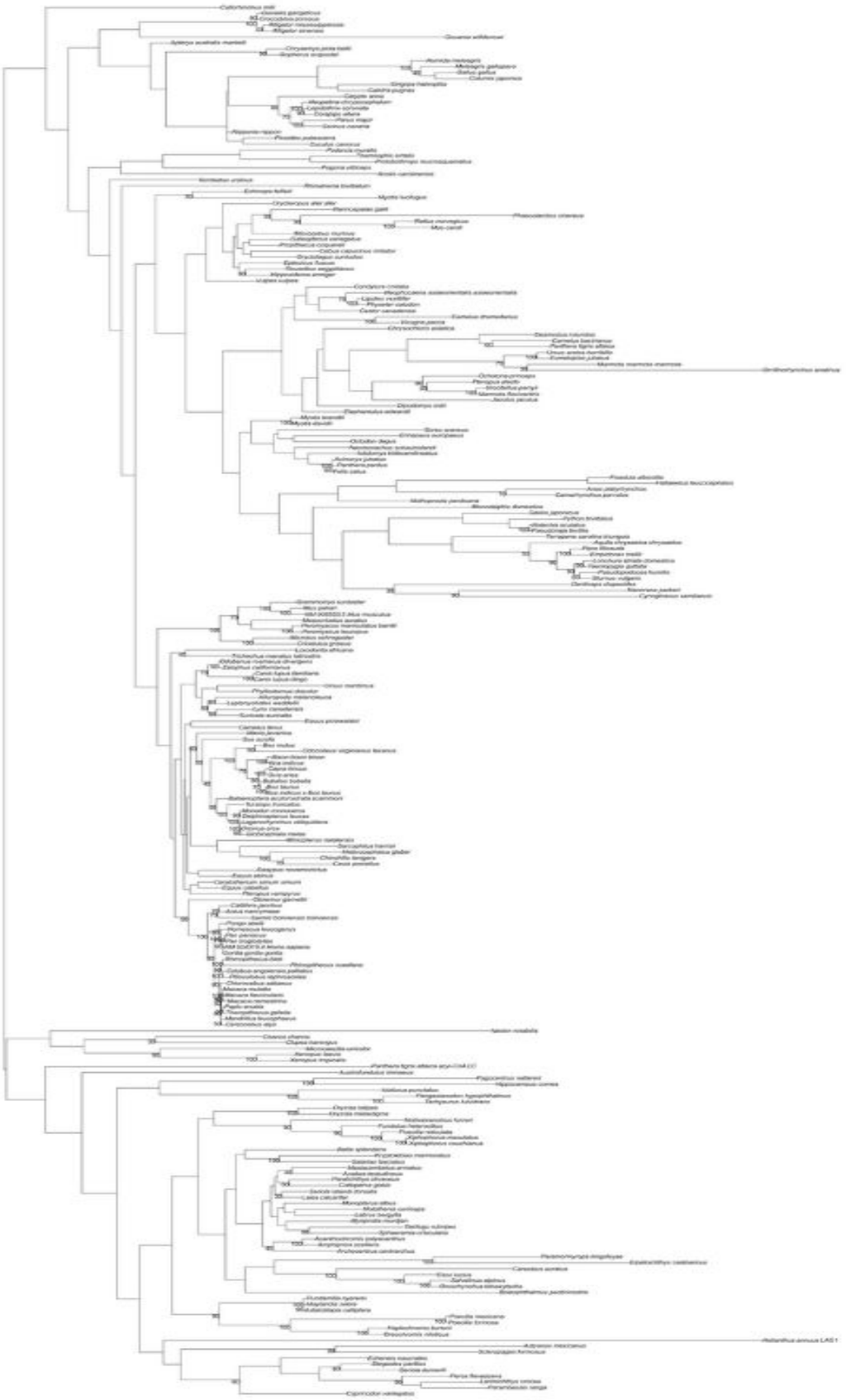


Fig 7. Decipher bootstrapping with JC model for 100 reps and 10 bootstrap trees. Annelida and Mammalia are well resolved. Full resolution image available here.

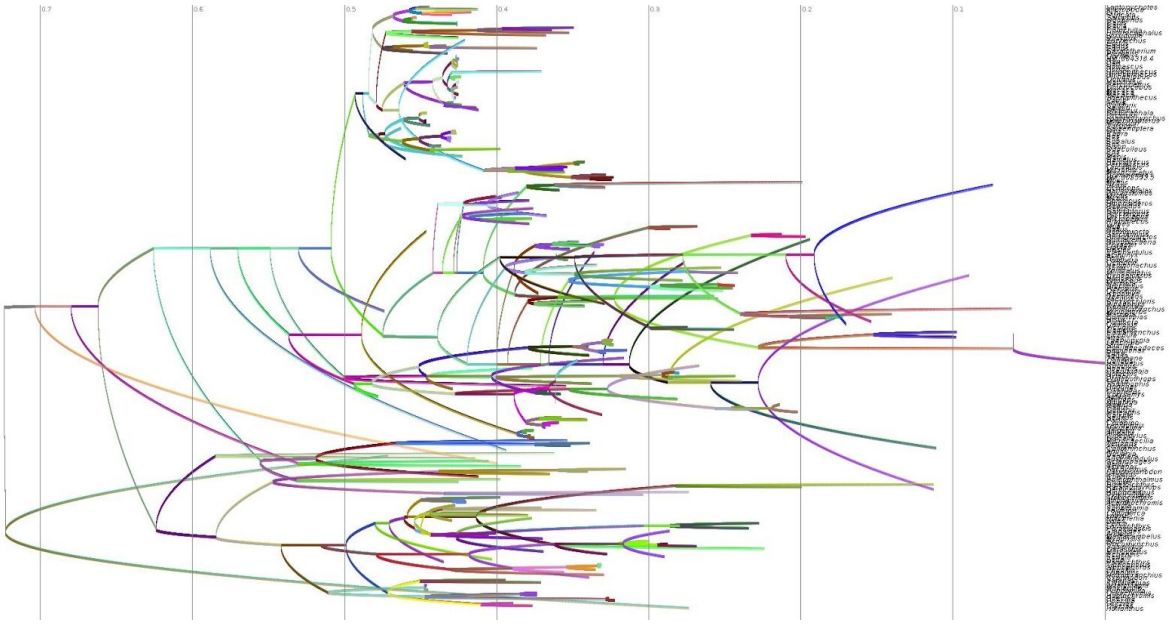


Fig 8. Consensus tree for 10 rounds of bootstrapping on Decipher alignment, and neighbor joining distance matrix. Density Tree software in java was used to produce this consensus image. Helianthus and other Annelids neatly group together in one big cluster. Additionally, Mammalian families form neat cultures, but fine resolutions isn't observed.

Tree produced by Decipher contains correct assignments for higher taxa. For Clustal Omega there are a lot of fine divisions and the grouping contains many artificial assignments.

Both algorithms show correct assignment of Mammalia as a distinct cluster from Annelida. The cluster for Annelida is completely treated as separate from all other clusters in case of Decipher, which also treats Helementhese as an outgroup albeit with many artificial assignments.

Figure 7, show the bootstrap tree generated from DECIPHER using 10 bootstraps, with Jukes Cantor model on the pml method with 100 repetitions. Figure 10, shows a consensus tree made in DensityTree software written in java. All images are available in repository for closer scrutiny.

Discussion

Dynamic programming approach in Decipher reduces a quadratic problem into linear time, resulting in rapid assignments. Additionally corrections for Homologues and Guide tree approach makes Decipher's performance comparable with Muscle. Clustal Omega produced a well resolved tree but with multiple artificial groupings, suggesting that it underperforms for a single tree. This can be compared with open tree of life data (McTavish, Holder, and Cranston 2018).

Additionally, time frames cannot be estimated as we have only a single gene as reference. Decipher treats this as a base and puts Helmenthese to far left of the tree, the distance is

suggestive of Helminths (Annelida) being the oldest of these organisms according to decipher. The fact that tree started from Ascl1 suggests that it is relatively newer addition to the evolutionary lot. Duplications and partial duplication of many genes, for example SRGAP2, segmental duplications, and so on have contributed towards an increased brain to size ratio for humans (Sassa 2013). It is suggestive that the same genes involved in primitive ganglion development have been spliced and reorganised to bring respective changes in higher organisms.

An interesting aspect is that Aves seems like an outgroup, even though it very well fits in the entire tree. Multiple species from Aves class are haphazardly assigned to different clades. It is noteworthy that Avian genomes are some of the smallest in the Animalia. Exploring this aspect could be interesting, since Avian genomes are relatively less cluttered.

It remains an open question, whether additional genes will give a better perspective. However, for a single gene the results produced by Decipher are promising in creating assessing primitiveness of species. It is suggested that Ascl1 and it's interactors may serve as guide for additional insights into advancement of mammalian, particularly Pan neo-cortex, and present biologically complicated questions like consciousness into simple modules.

References

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. "Evolutionary Dynamics of Language Systems." *Proceedings of the National Academy of Sciences of the United States of America* 114 (42): E8822–29.

"Interactome (protein Interactome)." 2015. *The Dictionary of Genomics, Transcriptomics and Proteomics*. https://doi.org/10.1002/9783527678679.dg06184.

McTavish, Emily Jane, Mark Holder, and Karen Cranston. 2018. "Nurturing a Sustainable Open Tree of Life." *Biodiversity Information Science and Standards*. https://doi.org/10.3897/biss.2.25727.

Sassa, Takayuki. 2013. "The Role of Human-Specific Gene Duplications during Brain Development and Evolution." *Journal of Neurogenetics* 27 (3): 86–96.

Sievers, Fabian, and Desmond G. Higgins. 2014. "Clustal Omega." *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi0313s48.

Wright, Erik S. 2015. "DECIPHER: Harnessing Local Sequence Context to Improve Protein Multiple Sequence Alignment." *BMC Bioinformatics*. https://doi.org/10.1186/s12859-015-0749-z.

https://simiacryptus.github.io/java-utilities/apidocs/com/simiacryptus/util/data/DensityTree.html