**Comparing Four Protein Homology Data Sources To Assess The Evolution Of TORC1 Complex In Plants. Implications For Conserved Domains Search.**

**Abstract**  
  
**Introduction**

In order to orchestrate a successful response to (biotic and abiotic) environmental changes, plants need to balance and coordinate their growth and stress responses.  
Autophagy is a subcellular recycling system conserved across all eukaryotes. The activation of autophagy is characterized by the formation of a double membrane structure called autophagosome. Inside the autophagosome, the cell deposits the different components (proteins, sugars, even whole organelles, etc.) to be degraded and its components reused by the cell. Autophagy has been shown to be crucial during developmental processes and stress response (Salem et al., 2018; Soto-Burgos et al., 2018; Wang et al., 2018). Even though autophagy activation and signaling pathway in plants is complex and still not completely understood, it has been shown that the process can be regulated by the Target of Rapamycin (TOR) complex (Soto-Burgos et al., 2018). In *Arabidopsis thaliana*, TOR regulatory complex is composed by the serine/threonine kinase TOR, the Regulatory Associated Protein of TOR (RAPTOR), and Lethal with Sec Thirteen 8 (LST8) (Soto-Burgos et al., 2018; Soto-Burgos and Bassham, 2017). Although TOR-independent regulation of autophagy has been reported (Pu et al., 2017; Soto-Burgos et al., 2018), we will focus on TOR-dependent regulation of the process for our work.

When environmental conditions are optimal, TOR complex (TORC) keeps autophagy and stress responses inhibited by phosphorylating a myriad of targets. One of these targets of phosphorylation is Autophagy Related 1 (ATG1) complex, known to be an important autophagy initiator (Avila-Ospina et al., 2014; Kamada et al., 2010; Lv et al., 2014; Soto-Burgos and Bassham, 2017). Upon sensing an environmental stress (such as nutrient starvation or drought stress), the energy sensor SNF1-related protein kinase (SnRK1) inhibits TORC (Soto-Burgos and Bassham, 2017). At the same time, SnRK1 can directly phosphorylate ATG1 to activate autophagy response (Avila-Ospina et al., 2014; Soto-Burgos et al., 2018; Soto-Burgos and Bassham, 2017).

Autophagy signaling and regulation pathway is known to crosstalk with different hormone signaling pathways to balance the plant growth/stress response. Upon drought stress sensing, RAPTOR is phosphorylated by SnRK2s, inhibiting TORC activity (Wang et al., 2018). It has been also shown that Rho-like GTPase 2 (ROP2) can activate TORC in response to auxin presence (Schepetilnikov et al., 2017). Moreover, brassinosteroid (BR) signaling has been shown to be regulated by selective autophagy. In this case the BRI1-EMS-SUPPRESSOR 1 (BES1) transcription factor, one the BR signaling master regulator is being degraded by selective autophagy when *A. thaliana* plants are subjected to drought stress, suggesting an interesting crosstalk between BRs and autophagy (Nolan et al., 2017). Most of the published work show RAPTOR as the main target for regulating TORC activity (Lv et al., 2014; Michaeli et al., 2016; Salem et al., 2018; Wang et al., 2018; Xiong and Sheen, 2015). For instance, phosphorylation on RAPTORB Ser897 seems to be responsible for TORC inhibition by ABA/drought stress in *A. thaliana* (Wang et al., 2018). Nevertheless, the regulation sites present in one species may not be present in another. It has been shown in mammalian cell lines that GSK3 kinase can phosphorylate RAPTOR on Ser859 to inhibit TORC. However, that region of mammalian RAPTOR is not present in plants (Stretton et al., 2015). And, in order to have a better idea of which of these phosphorylation target sites are present in our model of study *A. thaliana*, we first need to assess a good phylogenetic reconstruction of the TORC components.

One of the main problems when assessing phylogenetic reconstruction of TORC components is that there is more than one database of homolog proteins, some of them are curated databases and some of them are only computational predictions. The more curated they are, the less plant sequences are represented in the database. In this work we try to assess four different data sources for homolog proteins and evaluate which one of them is the most informative for us to use in a phosphorylation sites/domains conservation analysis in order to obtain candidate regulation sites for TORC in *A. thaliana*.

**Materials and methods**

Dataset

Protein sequence for RAPTOR1B (AT3G08850.1), TOR (AT1G50030.1) and LST8-1 (AT3G18140.1) were retrieved from The Arabidopsis Information Resource (TAIR, www.arabidopsis.org). Four different sets of homolog proteins for each accession were obtained by using the following procedures:

* The first homolog proteins list was obtained from the National Center for Biotechnology Information (NCBI) HomoloGene website (<https://www.ncbi.nlm.nih.gov/homologene/>).
* The second list of homolog proteins was obtained by using the Basic Local Alignment Search Tool (BLAST, (Altschul et al., 1990)) with default settings and retrieving the top 100 best scoring hits for each of the three proteins. The search was performed against the “refseq\_proteins” database and viridiplantae organism filter (taxid:33090).
* The third list was obtained from the Compara Database at the ENSEMBL Plants website (<http://plants.ensembl.org/info/website/ftp/index.html>, (Kersey et al., 2018))
* The final list of homologs was obtained from Phytozome version 12.1 (<https://phytozome.jgi.doe.gov>, (Goodstein et al., 2012)) and using the Araport11 annotation for *Arabidopsis thaliana* (Cheng et al., 2017).

Each dataset was retrieved in FASTA format.

Data analysis

Multiple sequence alignment (MSA) of each dataset was performed using MAFFT software version 7.123b (Katoh and Standley, 2013) with the `--auto` argument.

Phylogenetic analysis of each dataset was performed on the CIPRES Science Gateway server version 3.3 (Miller et al., 2010). Maximum-likelihood phylogenetic trees were constructed using RaxML version 8.2.10 (Stamatakis, 2014) using `-N autoMRE` argument for bootstrap. Each tree was constructed using either BLOSUM62 (Henikoff and Henikoff, 1992) or LG (Le and Gascuel, 2008) as the amino acid substitution model with empirical base frequencies (`+F` argument). Outgroups were selected as follows: For BLAST datasets the lowest blast scoring sequence was used as outgroup, for Phytozome datasets the sequence with lower homology score was used as outgroup, for ENSEMBL we used the provided tree to select the outgroup sequence. It was not possible to choose a good outgroup for the HomoloGene datasets.

Optimized model parameters for each RaxML “bestTree” were assessed using RaxML-NG (Kozlov et al., 2019) with the `--evaluate` function.

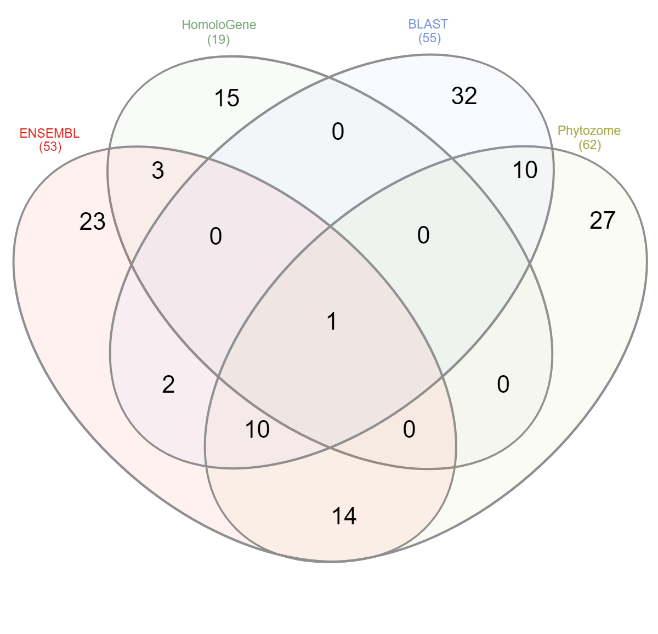
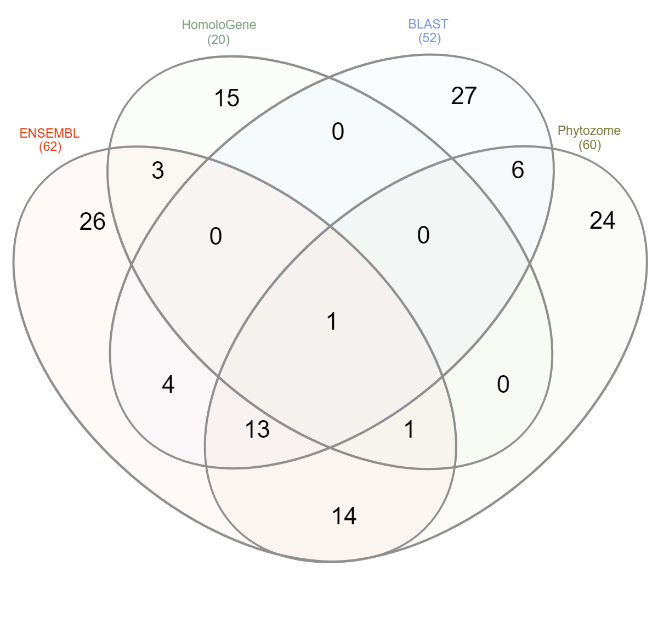
Resulting trees were visualized using FigTree version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

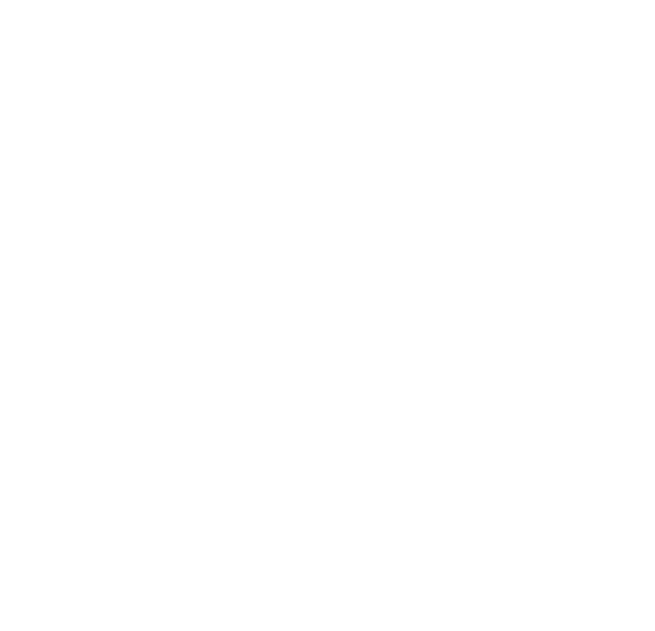
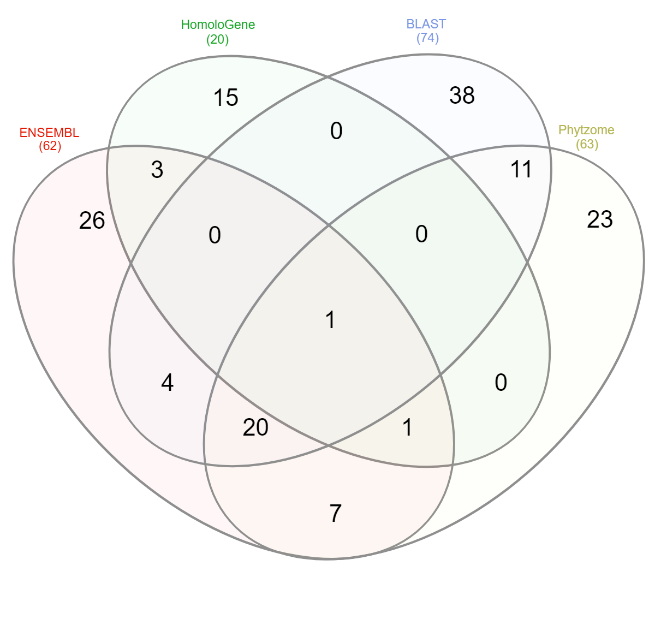
**Results and Discussion**

Different databases give different number of homolog sequences and species represented.

It was found that each queried database provided a different list of proteins when comparing the different datasets of homolog proteins for a specific protein. When searching for *A. thaliana* RAPTOR1B protein homologs, using BLAST gave us 100 sequences from 52 different species, using HomoloGene gave us 21 sequences from 20 different species, querying from ENSEMBL Plants gave us 116 sequences from 62 different species and, using Phytozome information gave us 101 sequences from 60 different species. For TOR protein, BLAST gave us 100 sequences from 56 different species, HomoloGene gave 19 sequences from 19 species, ENSEMBL Plants gave us 98 sequences from 53 different species, and Phytozome gave us 101 sequences from 62 different species. LST8-1 query gave 100 sequences from 76 different species using BLAST, 20 sequences from 20 different species in HomoloGene, 85 sequences from 62 different species using ENSEMBL Plants, and 101 sequences from 64 different species in Phytozome (Figure 1).

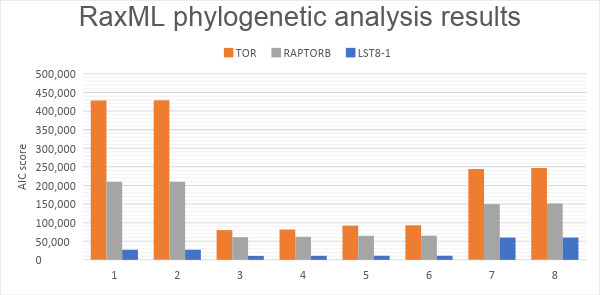






Phylogenetic analysis on different datasets reveals strong differences in likelihood.

Model testing revealed that, for every dataset, LG model gives higher tree probability than BLOSUM62 (Table 1). Maximum-likelihood (ML) phylogenetic analysis revealed a 3.5-fold difference in logLikelihood between datasets for RAPTOR1B, a 5.8-fold difference for LST8-1 datasets, and a 5.4-fold difference for TOR datasets (Table 1, Figure 2). This suggest that different available datasets of homolog proteins may provide different information for phylogenetic reconstruction.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 2 – RaxML phylogenetic analysis results for each dataset | | | | | |
| Gene | Database | Model | AIC | BIC | Final LogLikelihood |
| LST8-1 | ENSEMBL | LG | 27,166.084647 | 27,971.412736 | -13,397.042324 |
| LST8-1 | ENSEMBL | BLOSUM62 | 27,251.656081 | 28,056.984170 | -13,439.828041 |
| LST8-1 | NCBI\_BLAST | LG | 10,670.111351 | 11,531.184984 | -5,119.055676 |
| LST8-1 | NCBI\_BLAST | BLOSUM62 | 10,861.476567 | 11,722.550200 | -5,214.738283 |
| LST8-1 | NCBI\_HomoloGene | LG | 10,967.754516 | 11,191.833749 | -5,427.877258 |
| LST8-1 | NCBI\_HomoloGene | BLOSUM62 | 11,030.858383 | 11,254.937616 | -5,459.429191 |
| LST8-1 | Phytozome | LG | 59,912.367239 | 61,220.954566 | -29,738.183619 |
| LST8-1 | Phytozome | BLOSUM62 | 59,933.435596 | 61,242.022923 | -29,748.717798 |
| RAPTORB | ENSEMBL | LG | 209,942.242002 | 211,569.327155 | -104,723.121001 |
| RAPTORB | ENSEMBL | BLOSUM62 | 210,121.327868 | 211,748.413021 | -104,812.663934 |
| RAPTORB | NCBI\_BLAST | LG | 60,950.063336 | 62,107.433187 | -30,259.031668 |
| RAPTORB | NCBI\_BLAST | BLOSUM62 | 62,088.057614 | 63,245.427464 | -30,828.028807 |
| RAPTORB | NCBI\_HomoloGene | LG | 64,698.132707 | 65,027.930121 | -32,291.066354 |
| RAPTORB | NCBI\_HomoloGene | BLOSUM62 | 64,928.663384 | 65,258.460797 | -32,406.331692 |
| RAPTORB | Phytozome | LG | 149,858.931921 | 151,123.813552 | -74,711.465960 |
| RAPTORB | Phytozome | BLOSUM62 | 151,740.836203 | 153,005.717834 | -75,652.418101 |
| TOR | ENSEMBL | LG | 428,254.298283 | 429,832.087935 | -213,915.149142 |
| TOR | ENSEMBL | BLOSUM62 | 429,043.882293 | 430,621.671945 | -214,309.941146 |
| TOR | NCBI\_BLAST | LG | 79,872.396767 | 81,136.187423 | -39,720.198383 |
| TOR | NCBI\_BLAST | BLOSUM62 | 81,527.181821 | 82,790.972476 | -40,547.590910 |
| TOR | NCBI\_HomoloGene | LG | 92,140.823104 | 92,465.739902 | -46,016.411552 |
| TOR | NCBI\_HomoloGene | BLOSUM62 | 92,868.515928 | 93,193.432726 | -46,380.257964 |
| TOR | Phytozome | LG | 243,948.525709 | 245,413.692815 | -121,756.262854 |
| TOR | Phytozome | BLOSUM62 | 246,850.437068 | 248,315.604174 | -123,207.218534 |

Even though reconstructed trees differ between datasets, they give us a clear idea of A. thaliana closest relatives. When comparing the different phylogenetic trees obtained by using each dataset, we can appreciate a big difference between all of them (Figure 1, Supplemental trees figures). However, we were able to identify species that were always clustered in close relationship with A. thaliana, regardless of the dataset or gene used for the phylogenetic reconstruction. The one exception to this was the HomoloGene dataset which did not give us enough information about A. thaliana close relatives. This was mostly due to a lack of plant sequences on this database.

Code used:

species list

$ perl -ne 'if(/>.\*\_(.\*?)\//) {print $1."\n"}' RAPTORB\_ENSEMBL\_gene\_tree.fa > sp\_list.RAPTORB\_ENSEMBL\_gene\_tree.fa

$ perl -ne 'if(/(Org\_.\*?)\s/) {print $1."\n"}' RAPTORB\_phytozome\_top100.fasta > sp\_list.RAPTORB\_phytozome\_top100.fasta

$ gawk '{ if (match($0,/\[(.\*)\]/,m)) print m[0] }' RAPTORB\_NCBI\_BLAST\_viridiplantae100hits.fasta > sp\_list.RAPTORB\_NCBI\_BLAST\_viridiplantae100hits.fasta

Evaluate trees

$ raxml-ng --evaluate --msa ALIGNMENT\_FILE --model LG+F --tree bestTree.result --prefix GENE\_DATABASE\_MODEL

Fixing tree branch names

$ grep ">" RENAMED\_FASTA\_FILE > DATABASE\_old\_label.txt

$ perl -ne 'if(/ref\|(.\*?)\|.\*\[(.\*?)\]/) {print $1." ".$2."\n"}' GENE\_NCBI\_HomoloGene.fasta > homolo\_new\_label.txt

$ perl -ne 'if(/>(.\*?) .\*\[(.\*?)\]/) {print $1." ".$2."\n"}' GENE\_NCBI\_BLAST\_viridiplantae100hits.fasta > blast\_new\_label.txt

$ perl -ne 'if(/Org\_(.\*?) .\*peptide: (.\*?) /) {print $1." ".$2."\n"}' GENE\_phytozome\_top100.fasta > phyto\_new\_label.txt

Reformat the final tree

$ for FILE in \*.fixed;do perl -ne 'if(/^.\*?(\(.\*?;).\*/) {print $1}' $FILE > $FILE.nwk; done

Get AIC/BIC and logL values

$ find -name \*.raxml.log -exec grep -i 'AIC score' {} +  
$ find -name \*.raxml.log -exec grep 'Final LogLikelihood' {} +