**Comparing Four Main 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 grow and stress responses.  
Autophagy, a subcellular recycling system conserved across all eukaryotes, has been shown to be crucial during developmental processes and stress response.  
Brassinosteroids (BRs) are a steroidal plant hormone family known to be involved in a myriad of plant growth-related processes. Such as cell division/elongation and reproductive development.    
The GSK3-like kinase BIN2 (BRASSINOSTEROID-INSENSITIVE 2) is a key negative regulator of BR signaling response. This enzyme is known to phosphorylate the BR-responsive transcription factor BES1 (BRI1-EMS-SUPPRESSOR 1), preventing its nuclear accumulation and helping promote BES1 proteasomal degradation. Therefore, the plant BR response is inhibited.    
Broadening our knowledge about proteins being phosphorylated by BIN2 would give us a better picture of the different processes being controlled by this important regulator of BR response.   
  
**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 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).

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**

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 100 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 100 sequences from 63 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 100 sequences from 64 different species in Phytozome (Table 1).

Phylogenetic analysis on different datasets reveals strong differences in likelihood.

Model testing revealed that, for every dataset, LG model has higher tree probability than BLOSUM62 (Table 2). One exception is LST8-1 Phytozome dataset, which showed a slightly higher probability for the BLOSUM62 reconstructed tree.

Maximum-likelihood (ML) phylogenetic analysis revealed a 3.5-fold difference in logLikelihood between datasets for RAPTOR1B, a 5.5-fold difference for LST8-1, and a 5-fold difference for TOR (Table 2). This suggest that different available datasets of homolog proteins may provide different information for phylogenetic reconstruction.

Reconstructed trees

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

raxml-ng --evaluate --msa TOR\_phytozome\_top100.fasta.ren.align --model LG+F --tree RAxML\_\_\_\_\_\_\_\_\_\_\_\_23\_bestTree.result --prefix TOR\_phyto\_LG

grep ">" RAPTORB\_NCBI\_BLAST\_viridiplantae100hits\_renamed.fasta > blast\_old\_label.txt

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

perl -ne 'if(/>(.\*?) .\*\[(.\*?)\]/) {print $1." ".$2."\n"}' RAPTORB\_NCBI\_BLAST\_viridiplant

ae100hits.fasta > blast\_new\_label.txt

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

## Hypothesis  
By performing a phylogenetic analysis on the obtained sequence windows from our phospho-preoteomics dataset we can identify novel BIN2 target proteins.    
  
## Methodology    
  
  
1. Write a script to extract, from our dataset, the sequence window for each phosphorylated-site tagged as upregulated in response to BIN2 activity and to output these sequences as a multi-FASTA file.   
2. A multiple sequence alignment (MSA) will be performed on these sequences using either MUSCLE of MAFFT.    
3. The obtained MSA will be used to infer a phylogenetic tree (still don't know the method I will use for this)  
4. The obtained cluster o motifs will be used to further determine conserved sequence motifs, possible novel BIN2 targets and novel phosphorylation motifs.