**Phylogenetic Analysis of WRKY gene family in Maize**

**Yawei Li**

**EEOB 563: Molecular Phylogenetics**

**Dr.Dennis Lavrov**

**April 15th, 2021**

GitHub Link: https://github.com/YAWEIISME/EEOB563/tree/main/Final%20project

**Introduction**

The WRKY transcription factors constitute one of the largest transcription factor families in plants (Wang et al. 2018 ), which plays key roles in plant growth, plant development, defense regulation and stress response(He et al. 2016 ). WRKY transcription factors have been widely characterized in many plants, however, knowledge about this family in maize is limited.

In this study, I compared the 10 *WRKY*transcription factors, which were further classified into seven subgroups. Phylogenetic analysis showed that the 3 trees are constructed by multiple methods, including the neighbor-joining, maximum likelihood and the unweighted pair-group method with arithmetic mean (UPGMA) methods implemented in MEGAX on protein sequences of the WRKY domains are with relatively high accuracy and their results are matching.

**Research questions:**

There are several phylogenetic questions need to be solved by this study.

* What is the phylogeny for these transcription factors?
* How many groups and clades are contained in the built phylogenetic tree?

**Methods**

1. **Sequence data selection**

The WRKY transcription factors constitute one of the largest transcription factor families in plants, which plays key roles in plant growth, plant development, defense regulation and stress response. WRKY transcription factors have been widely characterized in many plants, however, knowledge about this family in maize is limited.

In this study, I choose the 10 *WRKY*transcription factors and download the protein sequence from the NCBI (National Center for Biotechnology Information).

### **Sequence Alignment**

**Multiple sequence alignment of the WRKY domains**

1. **Phylogenetic Analysis**

After sequences were aligned and configured for highest accuracy, phylogenetic trees were constructed by multiple methods, including the neighbor-joining (Tamura et al., 2011), maximum likelihood and the unweighted pair-group method with arithmetic mean

(UPGMA) methods implemented in MEGAX (<https://www.megasoftware.net/>) on protein sequences of the WRKY domains. Reliability of internal branches was assessed using the bootstrapping method (1,000 bootstrap replicates).

**Results**

1. **Sequence Alignment**
2. **Before alignment**

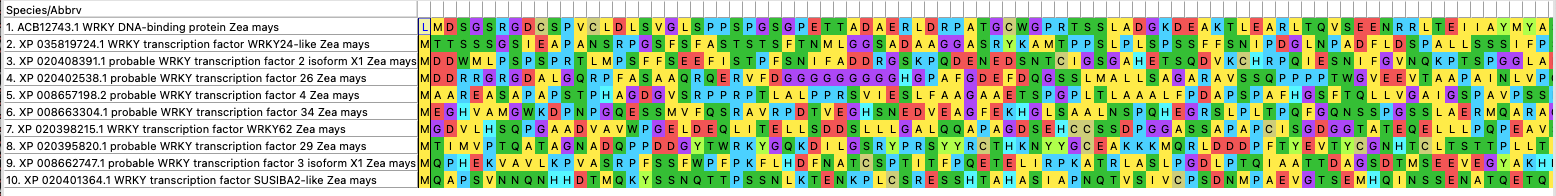


Figure 1: The multiple sequences before alignment

You can see the sequences with the different bases A-G-C-T are marked with different colors. These sequences are still arranged "chaotic" and not aligned neatly before the alignment.

1. **After alignment**

A screenshot of a computer

Description automatically generated with low confidence

Figure 2: The multiple sequences after alignment

There are many methods can be used to multiple sequences alignment. Here I choose the ClustalW, which is one of the most widely used [computer programs](https://en.wikipedia.org/wiki/Computer_program) used in [Bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics) for [multiple sequence alignment](https://en.wikipedia.org/wiki/Multiple_sequence_alignment). The gap opening penalty is 15.00 and the gap extension penalty is set as 6.66 as the default setting parameter. After the alignment is complete, we can see that the sequence alignments are neat with some gaps. Then I can start to construct a series of phylogenetic trees.

1. **Phylogenetic Analysis**
2. **Neighbor-joining (NJ-tree)**

Text

Description automatically generated

Figure 3: The unrooted Neighbor-joining (NJ-tree)

The value on the neighbor-joining tree shows the distance between each pair of [taxa](https://en.wikipedia.org/wiki/Taxa). We can see that XP 008657198.2 probable WRKY transcription factor 4 Zea mays and XP 020402538.1 probable WRKY transcription factor 26 Zea mays are the most closely related, while ACB12743.1 WRKY DNA-binding protein Zea mays is relatively far with other groups.

1. **Maximum likelihood-MML tree (Original tree)**Text

   Description automatically generated

Figure 4: Maximum likelihood-MML tree (Original tree)

1. **Maximum likelihood-MML tree (Bootstrap Consensus Tree)**

Table

Description automatically generated

Figure 5: Maximum likelihood-MML tree (Bootstrap Consensus Tree)

Interestingly, we can see from the Maximum likelihood tree that the possibility of the clustering of the XP 020395820.1 probable WRKY transcription factor 29 Zea mays and XP 020398215.1 WRKY transcription factor WRKY62 Zea mays is higher than that of XP 008657198.2 probable WRKY transcription factor 4 Zea mays and XP 020402538.1 probable WRKY transcription factor 26 Zea mays. But the ACB12743.1 WRKY DNA-binding protein Zea mays still shows the relatively far distance with other groups.

1. **UPGMA-tree (original tree)**

Text

Description automatically generated

Figure 6: UPGMA-tree (Original tree)

1. **UPGMA-tree (Bootstrap Consensus Tree)**

Table

Description automatically generated

Figure 7: UPGMA-tree (Bootstrap Consensus Tree)

The results from UPGMA-tree are different with the other two models and the XP 008657198.2 probable WRKY transcription factor 4 Zea mays and XP 020402538.1 probable WRKY transcription factor 26 Zea mays are clustered into different subgroups.

There are 2 types of trees for each model, including the Original Tree and the Bootstrap consensus tree (the tree merged by the step size test).

**Discussion:**

Test of Phylogeny (testing method of establishment) is used to test the quality of establishment. The default test method is Bootstrap method (step test). Step length inspection needs to set the number of inspections, usually a multiple of 100, the default setting is 500, I choose 1000 in my study.

The step length test is to calculate and draw the specified number of phylogenetic trees according to the selected tree-building method. Because the core algorithm of most tree-building methods is a statistical probability model, the tree calculated each time will be different.

Each node on the established phylogenetic tree will be marked with a number, which represents how many percent of the phylogenetic tree calculated by the specified number of times contains this node. Generally, trees whose values ​​on most nodes are greater than 70% can be trusted. Individual nodes below 70% can be tolerated temporarily, or the quality can be improved by adding or mountain sequences.

Substitution Model. It is the calculation model used when selecting the genetic distance. In theory, we should try various models before selecting the most suitable model calculation according to the test results. But here I choose a simpler distance model, such as p-distance.

Gap/Missing Data Treatment, most of the establishment methods will require the deletion of columns with more gaps in the multiple sequence alignment. But according to the different genetic distance measurement methods, the deletion principle is also different. If the genetic distance is measured by the number of different residues between the sequences, the NJ nearest neighbor method used here can be Partial deletion. The degree of deletion is set at 50%, that is, half of the columns with gaps are retained. If it is another method of establishment, you need to select Complete deletion.

There are 2 types of trees for each model, including the Original Tree and the Bootstrap consensus tree (the tree merged by the step size test). On the Bootstrap consensus tree, the number at the node shows that the percent of the tree has this branch, which indicates the credibility of the branch after the step length test. In the currently constructed phylogenetic tree is generally credible based on the value at most of the nodes are ≥70. The original Tree is one of the 1,000 trees constructed by the step length test, which is not merged by multiple trees. So the length of the branches can accurately represent the genetic distance. In addition, it can also be seen from this tree whether the previous artificial grouping situation has undergone unexpected changes. For example, some may seem to break away from the grouping and become an outer group, thus determining the root of the tree.

**References:**

He, G. H., Xu, J. Y., Wang, Y. X., Liu, J. M., Li, P. S., Chen, M., Ma, Y. Z., & Xu, Z. S. (2016).

Drought-responsive WRKY transcription factor genes TaWRKY1 and TaWRKY33 from wheat confer drought and/or heat resistance in Arabidopsis. BMC Plant Biol. 2016 May 23;16(1):116.doi: 10.1186/s12870-016-0806-4.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Wang, C. T., Ru, J. N., Liu, Y. W., Li, M., Zhao, D., Yang, J. F., Fu, J. D., & Xu, Z. S. (2018). Maize WRKY Transcription Factor ZmWRKY106 Confers Drought and Heat Tolerance in Transgenic Plants. *International journal of molecular sciences*, *19*(10), 3046. https://doi.org/10.3390/ijms19103046