FMO SUPERFAMILY PROTEIN PHYLOGENY AND THE ORIGIN OF YUCCA FAMILY.

Turnaev I.I.[1], Suslov V.V.[1], Gunbin K.V.[1], Afonnikov D.A. [1].*

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia.*
* *Corresponding author e-mail:* turn@bionet.nsc.ru

Motivation and Aim: The evolutionary origin of the YUCCA family is currently under discussion. We conducted a comparative analysis of the class b flavin-containing monooxygenase family proteins to clarify the origin of YUCCA.

Results and Discussion: We constructed a phylogenetic tree for class B flavoprotein monooxygenases poteins, including protein and transcriptome sequences, using the IQ-TREE program - "FMO_prot&transcr-IQ-TREE" (fig.1A). Class G FMOs proteins were selected as an outgroup. Class B flavoprotein monooxygenase poteins include three subclasses: {1-5, 7-8 protein groups} FMOs, {6} NMOs, and {9}BVMOs (fig. 1A). On the tree, we identified 8 groups, one of which is type II FMO divided into three subgroups: {4} type IIa FMO, {5} type IIb FMO, and {7} type IIc FMO (see fig. 1A for group names and numbers).

Further, a detailed comparative analysis of conservative sites and the domain composition of proteins of these groups was carried out. As a result, it turned out that the sequences of {5} type IIb FMO (fig.1A,B) differ significantly from other groups of proteins in the FMOs subclass in the composition of conservative sites, domain composition, and belonging to taxa – this group is represented in plant, fungal, and bacterial proteins – in contrast to only bacteria proteins in {4} type IIa FMO, {5} type IIb FMO, {3} YUC-like bacteria, {2} cyanobacteria FMO, and only lend plants in {1} YUCCAs.

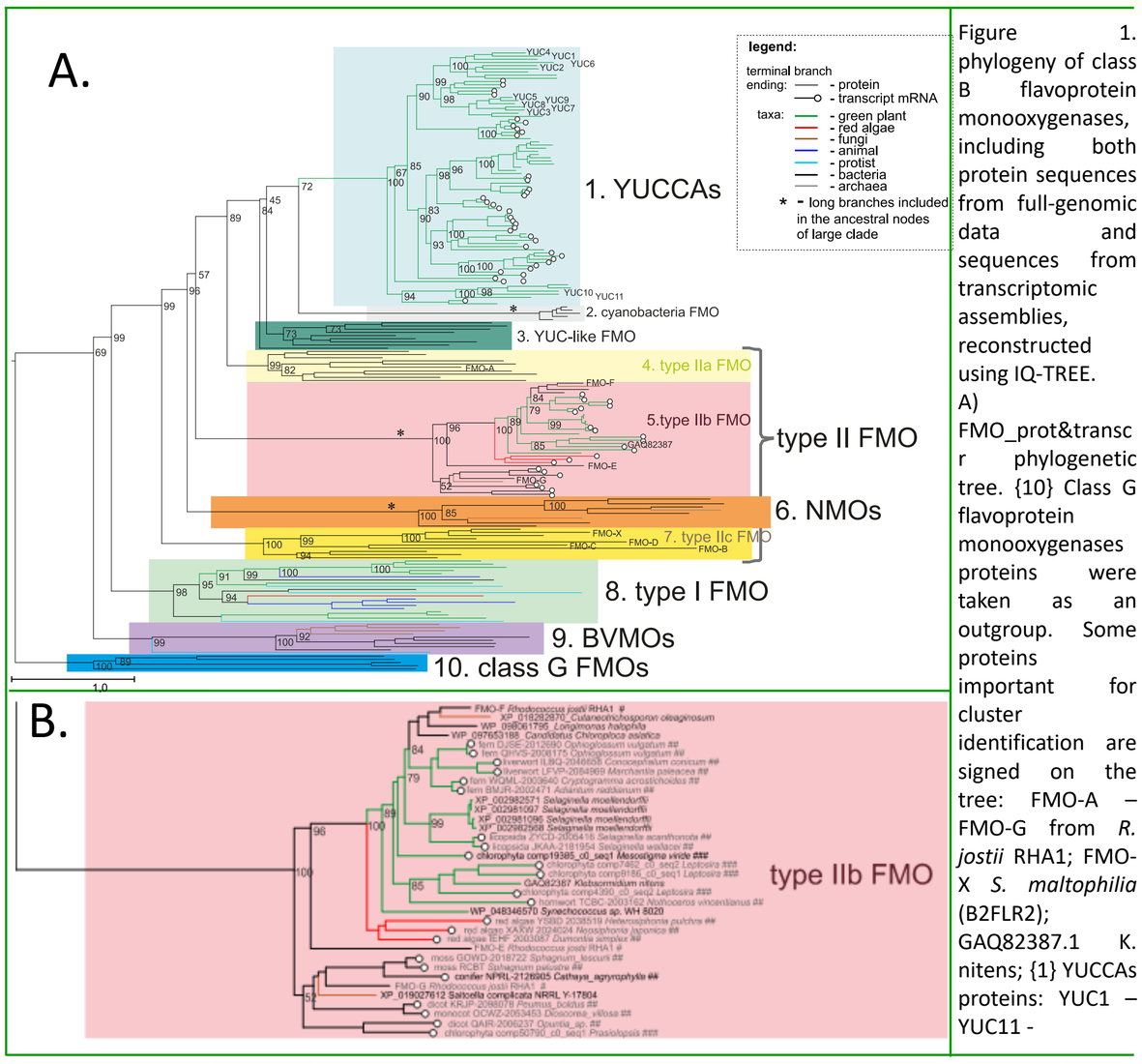

Figure 1. phylogeny of class B flavoprotein monooxygenases, including both protein sequences from full-genomic data and sequences from transcriptomic assemblies, reconstructed using IQ-TREE. A) FMO_prot&transcr phylogenetic tree. {10} Class G flavoprotein monooxygenases proteins were taken as an outgroup. Some proteins important for cluster identification are signed on the tree: FMO-A – FMO-G from *R. jostii* RHA1; FMO-X *S. maltophilia* (B2FLR2); GAQ82387.1 K. nitens; {1} YUCCAs proteins: YUC1 – YUC11 - *A. thaliana* proteins (gene IDs from Plaza plant database for these YUCCAs genes are shown in the caption to fig. 1A. B) Fragment of the FMO_prot&transcr tree (fig. 1A), including only the {5} type IIb FMO clade. The designations are similar to those in figure 4A, in addition, the names of protein sequences from NCBI and taken from the article Riebel et al., 2013 are highlighted in black text; transcriptomic sequences are highlighted in gray text. The three protein sequences marked with the # icon are taken from the Riebel et. al. study, 2013; # # - taken from the 1000 plant genomes database (https://sites.google.com/a/ualberta.ca/onekp/) and # # # - from the Green Algal Tree of Life project [https://algae.eeb.uconn.edu/research-interests/green-algae-tree-of-life-project/].

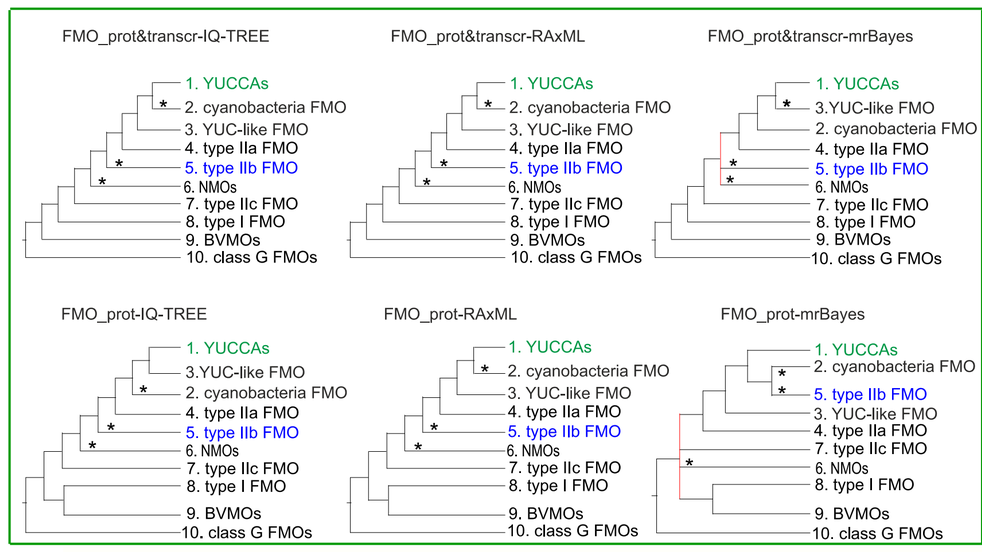

Figure 2. Simplified cladograms of class B flavoprotein monooxygenases constructed using three methods: IQ-TREE (phylogenetic tree in fig. 1A), RAxML and mrBayes). The upper row – cladograms for FMO_prot&transcr alignment; lower row - cladograms for FMO_prot alignment. {1} The YUCCAs clade is highlighted in green, and {5} the type IIb FMO clade is highlighted in blue. * - long branches included in the ancestral nodes of large clades on the tree. The red line marks a trifurcation in the cladogram reconstructed by mrBayes.

To assess the stability of the phylogeny in addition to reconstructing the phylogeny of FMO_prot&transcr-IQ-TREE (fig. 1A and its simplified cladogram-fig. 2, upper row) a similar reconstructions was performed using RAxML (FMO_prot&transcr-RAxML) (only the simplified cladogram is presented – fig. 2, upper row) and mrBayes (FMO_prot&transcr-mrBayes) (only the simplified cladogram is presented – fig. 2, upper row).

And figure 2 (bottom row) shows three analogous cladograms for sampling without transcriptome sequences – FMO_prot-IQ-TREE, FMO_prot-RAxML, and FMO_prot-mrBayes.

The topology of FMO_prot&transcr-IQ-TREE, FMO_prot&transcr-RAxML, FMO_prot-IQ-TREE, FMO_prot-IQ-TREE trees is conservative. However, the topologies of FMO_prot-mrBayes to a greater extent than FMO_prot&transcr-mrBayes have important differences from the topology of the aforesaid trees: 1) the FMO_prot-mrBayes and FMO_prot&transcr-mrBayes trees are not up to resolved – they have tetrafurcation and trifurcation, respectively; 2) in FMO_prot-mrBayes, the {5} type IIb FMO clade (highlighted in blue text to fig. 2) approaches with {1} YUCCAs, so they become sister. so that they become sisterly.

This topology shift can be caused by the LBA effect, since the trees have three long branches that entering in ancestral nodes of the following clades: {2} cyanobacteria FMO, {5} type IIb FMO, and {6} NMOs (in fig. 1A, fig. 2 these long branches are marked *). Based on a comparative analysis of the phylogenies that we received, we assume that the Bayesian method, the mrBayes program, was more sensitive to LBA effect than the Maximum Likelihood method (implemented in IQ-TREE and RaxML programs), in our stady. Our assumption is consistent with the conclusions of Kolaczkowski and Thornton [Kolaczkowski and Thornton, 2009], which showed that the Bayesian inference (BI) of phylogenetic relationships, in contrast to Maximum Likelihood methods (ML), leans in favor of topologies that group long branches together, which leads to an artifact topology of trees.

Thus, we tend to favor the phylogeny topology of class B flavoprotein monooxygenases, where the clade {5} type IIb FMO is remote from the clade {1} YUCCAs (fig. 1C, 4C), since it is found in all trees except FMO_prot-mrBayes (fig. 1C). In addition, only trees built by mrBayes are under-resolved.

There are also significant differences in the cosenses of FMO identified (FxGxxxHxxxH[70%]/y[20%]/f[10%]) and FAD binding (GxN[75%]/g[17%]/c[13%]xxA/G) sites for {5} type IIb FMO proteins compared to FMO identified (FxGxxxHxxxY/F/W) and FAD binding (GxGxxA/G) sites for the rest of the class B flavoprotein monooxygenases poteins (groups 1-4, 6-9 in fig. 1A).

Conclusion: 1) We have identified a new group of type IIb FMO proteins, including plant proteins (among them Charophyta GAQ82387. 1 proteins from *K. nitens* described earlier by Wang et al., 2014) and FMOe, FMOf, FMOg bacteria proteins from *R. jostii* RHA1 (described earlier by Riebel et al., 2013).

2) Our data suggest that the type IIb FMO and YUCCAs proteins occurred in evolution independently of two different HGTs. This supports the hypothesis of Yue et al. about the origin of YUCCAs as a result of HGT from bacteria to land plants, but not the hypothesis of Wang et al. about the appearance of YUCCAs in evolution already in Charophyta.

Facts from literary sources:

I) Riedel et. al. (Riedel et. al., 2013) in a study of actinobacteria *Rhodococcus jostii* RHA1 proteins identified subclass from the class B flavoprotein monooxygenases, which they named type II FMOs. Of the eight proteins (FMS ag proteins) identified by them in *R. jostii* RHA1, three (FMOe, FMOf, FMOg) had special catalytic properties not typical for FMOs subclass proteins. They were able to catalyze not only the reaction of sulfoxidations (the ability of FMOs and BVMOs), but also Baeyer–Villiger oxidations (the ability of BVMOs). In addition, FMOe, FMOf, FMOg proteins can use both NADPH and, much cheaper, NADH as a coenzyme. Although all other FMOs proteins use only NADPH . This makes these three proteins promising targets for their use in biotechnology [Riebel et al., 2013].

II) Based on their research, Yue. et al. (2014) proposed a hypothesis about the origin of {1} yuccas in the most recent common ancestor (MRSA) of the main land plant taxa as a result of HGT (horizontal gene transfer) from bacteria. Later, Wang et al. (Wang et al., 2015) based on the presence of {1} YUCCAs homologs of the GAQ82387.1 protein in Charophyta *Klebsormidium nitens*, it was suggested that YUCCAs proteins appeared in evolution already in Charophyta (Fig. 1A, GAQ82387.1 protein).

Methods and Algorithms: Recognition of sequences for forming samples was performed by the BLASTP program. Multiple sequence alignment was performed using Promals and Mafft. IQ-TREE 1.6.12, RAxML 8.1.24, and mrBayes 3.2.7 programs were used for phylogenetic tree reconstruction.

Acknowledgements: The work was supported by project 0324-2019-0040.

References:

1. Riebel A., de Gonzalo G., FraaijeIn M. F., Expanding the biocatalytic toolbox of flavoprotein monooxygenases from Rhodococcus jostii RHA1. *Enzymatic*, 2013, 88, 20– 25.
2. Wang C., Liu Y., Li S.-S., and Han G.-Z., Origin of plant auxin biosynthesis in charophyte algae. *Trends in Plant Science*, Vol. 19, No. 12 pp. 741-743.
3. Yue J., Hu X., and Huang J. Origin of plant auxin biosynthesis. *Trends in Plant Science*, Vol. 19, No. 12 pp. 764-770.
4. Kolaczkowski B., Thornton J.W.. Long-Branch Attraction Bias and Inconsistency in Bayesian Phylogenetics. PLoS One., 2009, 4(12), e7891.