



**S Fig 2: HMMERCTTER Analysis of the Alpha Crystallin Domain Protein Superfamily.**

Numerical analysis in Table 1, details in context. Colors according to HMMERCTTER output, leaves in black could not be clustered. Scale bars indicate 0.1 amino acid substitution per site. (A) Detail of the classification upon incorrect clustering (Run 1 with with classes indicated numerically) of sequence VV00193000. Classified sequences of R1\_11 in red; R1\_15 in gray and R1\_22 in light green lines. Note that R1\_15 is nested into the clade containing R1\_22. Target sequence ME22590va contains three ACDs also generating classification conflicts. (B) Detail of final classification Run 2 obtained upon removal of VV00193000 and ME22590va from training and target-set, respectively. (C) Training Tree with clusters numbered according to HMMERCTTER and codes applied by Bondino et al., (M+ combines mitochondrial (M) and mitochondrial-like sHSPs (ML)) note that UAPVII is represented by four clusters. UAPI and II were originally identified in the final 17 proteome dataset and correspond to a single clade in the training tree. I indicates a single orphan sequence, II points to a an example of change in tree topology as compared with (D), the final and complete reference tree. (E) Detail of cluster 7/M+ demonstrating that particularly sequences at large distances (ML) are often not detected. Sequences with purple boxed labels are training sequences. The arrow points to a monocotyledon sub-clade.