Phylogenetic Analysis of Plant Transcription Factor Gene Families

Daniel Lang*, Sandra Richardt, Ralf Reski, Wolfgang Frank, Stefan A. Rensing

Plant Biotechnology, Faculty of Biology, University of Freiburg,
Schaenzlestr. 1, 79104 Freiburg, Germany www.plant-biotech.net; www.cosmoss.org
*daniel.lang@biologie.uni-freiburg.de; phone: +49 761 203 6974; fax: +49 761 203 6990

Transcription factors (TF) regulate biological processes via control of gene expression. The classification into families is based on their highly conserved DNA-binding domain. In the course of evolution, an increased rate of duplicate retention and domain shuffling gave rise to a significant expansion and high diversity of TF families in plants, many of which are not found in other eukaryotes (e.g. HD-Zip, MADS MIKC-type).
During the last years, TF genes were identified in the seed plants *Arabidopsis* and rice and the distribution of TF families was compared with each other as well as with other eukaryotes. Additionally, phylogenies of several single TF families were characterised among plants (e.g. AP2, MADS, Sigma factor, FLO/LFY). However, a global phylogenetic analysis of plant TF families spanning the whole green lineage and in particularl including non-seed plants, i.e. mosses and algae, was missing until now.
In terms of evolution, mosses are located halfway between seed plants and algae. Therefore, we used the moss *Physcomitrella patens* as an offset for the phylogenetic analyses to globally analyse genes of plant TF families. On the basis of the comprehensive *Physcomitrella* transcriptome (www.cosmoss.org) we identified, functionally annotated and classified more than 1,000 putative TF sequences by BLAST, pattern and HMMer searches. To facilitate phylogenetic analyses of this large dataset we have developed the TreePipe, an automated phylogenetic pipeline combining state of the art methods like PSI-BLAST, ProbCons, Muscle, MAFFT, TreePuzzle and several clustering algorithms to construct phylogenetic trees for large datasets without manual interference. Additionally, we developed a new method, the Taxonomic Profile, in order to infer the taxonomic distribution of cluster members from the NCBI taxonomy information and subsequently use it for average linkage clustering and heap map visualisation. This provides a tool for easy identification of TF families with correlating or biased taxonomic distribution.
The automated procedure yielded putative TF gene clusters for which known domain structures were determined using InterProScan. Manual annotation resulted in 180 TF clusters covering all known major families, 70 clusters of other transcriptional regulators, and 50 clusters of so far unknown function. The full information gathered throughout the automated process encompassing functional, taxonomic and domain structure annotation as well as the corresponding phylogenetic trees were combined using data mining techniques and manual annotation, leading to a comprehensive resource of plant transcription factors, which will be made available soon.

Financial support by the DFG (Re 837/10-1) ist gratefully acknowledged.