Plant Transcription Factors @ uni-potsdam.de

Diego Mauricio Riaño-Pachón, Ingo Dreyer, Slobodan Ruzicic, Bernd Mueller-Roeber
Department of Molecular Biology, Institute for Biochemistry and Biology, University of Potsdam†
email: {diriano,dreyer,ruzicic,bmr}@uni-potsdam.de

Introduction

Transcription factors are proteins that regulate gene expression levels by binding to specific short DNA sequences (*cis*-acting elements) in the promoter region of target genes, and enhancing or repressing their transcription rate. Their identification is a key step for the reconstruction of transcriptional regulatory networks.

The availability of complete genomes together with protein domain models (e. g., HMMS) in public repositories (e. g., PFAM, SMART) allows to predict sets of putative transcription factors on a genome-wide scale in several species (e. g., [1, 2]). Here we have predicted the putative complete set of transcription factors in five plant species.

Methods

The first step was the collection of information from the literature about transcription factor families in plants (e. g. [3]). We used this information to devise a set of rules based on the presence of certain protein domains, mainly specific DNA-binding domains, to group the complete set of proteins of individual organisms into up to 68 different families. The search for protein domains was carried out using HMMER v2.3.2 and PFAM v20.0. Sixty-one PFAM domains were employed in the classification scheme (Fig. 1). In a few cases no appropriate domain model was found in PFAM, consequently we created our own profile-HMMs based on published multiple alignments, or through PSI-BLAST searches (i. e. Alfin-like, CCAAT-Dr1, DNC, G2-like, GRF, HRT, LUFS, NF-YB, NF-YC, NOZZLE, STER_AP, trihelix, ULT, VOZ).

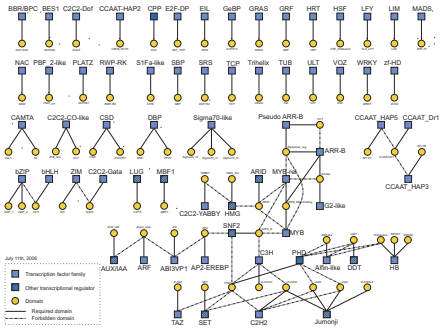

Figure 1: Graph representing the relationships between transcription factor families and protein domains.

Rules to assign membership to TFs families are represented as a bipartite graph (Fig. 1), where the blue nodes represent protein families (i. e. transcription factors or other transcriptional regulators (shaded)) and the yellow ones represent protein domains. Two types of edges are shown, a continuous edge represents a required relationship, i. e. a domain must be present in a protein to be assigned to a family. In a similar way, discontinuous edges represent a forbidden relationship, where the definition of a family excludes the presence of a given domain. For 31 families the presence of a single domain was enough to assign membership, the remaining families needed combinations of domains.

Total number of transcription factors

We have applied this approach to the complete proteome of five plant species, whose genome is in an advanced process of annotation, i. e. two green algae and three angiosperms: *Ostreococcus tauri* (University of Ghent), *Chlamydomonas reinhardtii* (Joint Genome Institute/US Department of Energy), *Arabidopsis thaliana* (The Arabidopsis Information Resource), *Populus trichocarpa* (JGI/DOE), *Oryza sativa* (The Institute for Genomic Research).

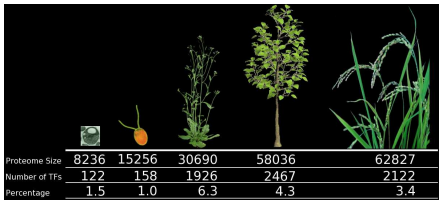

Figure 2: Transcription factors per species.

Transcription factor families

On the one hand, the total number of observed transcription factors per genome (Fig. 2) seems to follow a power-law , as it has been stated before (e. g. [4]). However, the number of data points is too small to have reliable estimates for the fit of the parameters of a power-law. On the other hand, the number of TFs per family varies widely (Fig. 3).

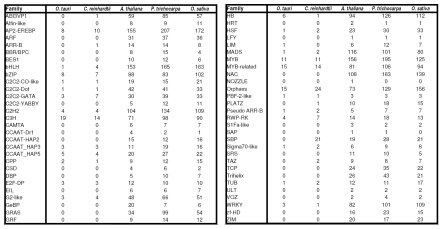

Figure 3: Number of TFs per family

Despite the differences between absolute numbers of TFs per family, it can be seen in figure 4 and mainly in fig 5 that the three angiosperms vary with the same tendency, and the same occurs for the two green algae. When green algae and angiosperms are compared the relation is less strong but still a high proportion of the variance ($\sim 40\%$) in one can be explained by the variability in the other.

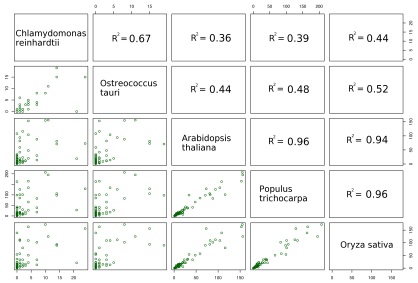

Figure 4: Relation between the numbers of member genes per TF family in different species.

It is clear that TF families have been differentially expanded in plants (Fig. 5). But the most abundant families (after normalization) are the same in all species, i. e. C3H, MYB, AP2-EREBP, bZIP and C2H2, additionally bHLH, NAC, MADS, HB, WRKY, ABI3VP1 are common among angiosperms.

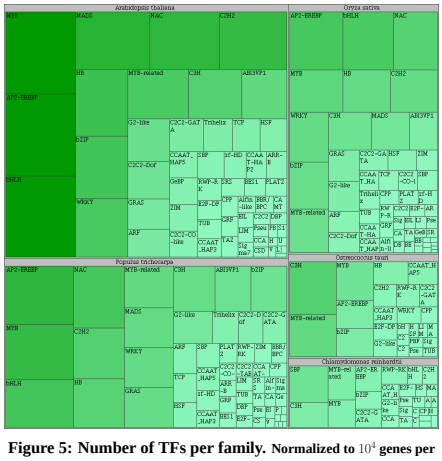

Figure 5: Number of TFs per family. Normalized to 10^4 genes per genome

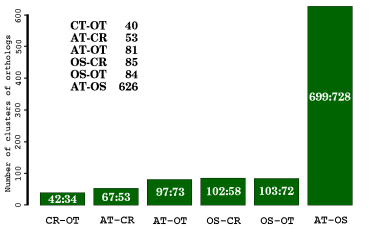

Figure 6: Pairwise number of clusters of orthologs

Most of the observed TFs appeared after the divergence between green algae and angiosperms (Fig. 6). They could have appeared by domain shuffling, sharing little sequence conservation with other sequences. To have a better understanding about the evolutionary process leading to uncreasing numbers of TFs we would need the genome annotation of bryophytes, and gymnosperms. Genome sequencing of representative species is currently ongoing.

Web databases

Information about transcription factor families and their member genes is being stored in a relational database powered by MySQL and made publicly available (*O. tauri*, *C. reinhardtii*, *A. thaliana* and *O. sativa*) on the WWW:

`http://ostreotfdb.bio.uni-potsdam.de`
`http://chlamytfdb.bio.uni-potsdam.de`
`http://arabtfdb.bio.uni-potsdam.de`
`http://ricetfdb.bio.uni-potsdam.de`

The databases are cross-referenced by means of ortholog relationships between pairs of species, determined using the INPARANOID software [5].

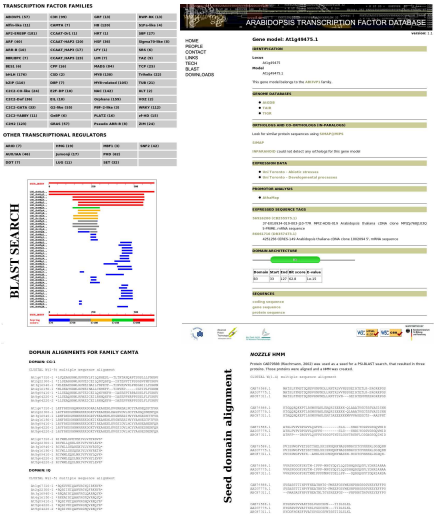

Figure 7: Screen shots.

For each family we have included a short description and relevant references, with links to NCBI's PUBMED. There are domain alignments for the family members, and each putative TF has a detailed information page (Fig. 7).

We also provide information about the domain for which we created profile-HMMs and the information employed in the definition of TF families.

Future

Additional species will be included, focusing in bridging the gap that we have now between green algae and angiosperms.

We will give access to users from the TF community to upload annotations/comments into the database for individual genes.

Acknowledgements

This work was supported by the European Union (NICIP; EU CT-2002-00245) and the Center for "Advanced Protein Technologies" of the University of Potsdam.

References

[1] Kummerfeld, S. K. and Teichmann, S. A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res*, **34**, D74–D81.

[2] Davuluri, R. V., Sun, H., Palaniswamy, S. K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

[3] Riechmann, J. L., et al. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

[4] van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends Genet*, **19**, 479–484.

[5] Remm, M., Storm, C., and Sonnhammer, E. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, **314**, 1041–52.

†Address for correspondence: Universität Potsdam, Institut für Biochemie und Biologie, Karl-Liebknecht-Str. 24-25, Haus 20, 14476 Golm, Deutschland