

Poster M-4

Plant Transcription Factors



Authors:

Diego Mauricio Riano Pachón (*Department of Molecular Biology, Institute of Biochemistry and Biology, University of Potsdam*)

Ingo Dreyer (*Department of Molecular Biology, Institute of Biochemistry and Biology, University of Potsdam*)

Slobodan Ruzicic (*Cooperative Research Group, Max Plack Institute of Molecular Plant Physiology*)

Bernd Mueller-Roeber (*Department of Molecular Biology, Institute of Biochemistry and Biology, University of Potsdam*)

Short Abstract: Transcription factors (TF) are proteins that play a central role in the regulation of gene expression; usually they are members of gene families. Here we have identified the members of up to 53 TF families of plants whose genome sequences are available (*Oryza sativa*, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Ostreococcus tauri*).

Long Abstract:

Plant Transcription Factors <p>A list of protein domains involved in plant transcriptional regulation was built through keyword searches at the PFAM's website and from literature. PFAM (Bateman et al. 2004) global profile-HMMs for those domains, together with HMMER package (Eddy 1998), were employed to search the complete protein set from <i>Oryza sativa</i> (latest genome annotation from TIGR, 2006), <i>Arabidopsis thaliana</i> (TIGR, 2004), <i>Chlamydomonas reinhardtii</i> (JGI/DOE, 2006) and <i>Ostreococcus taur</i>i (University of Gent, 2006). In a few cases, when no appropriate domain models were found in PFAM, we created our own profile-HMMs (e. g. G2-like, NF-YB, NF-YC, CCAAT-Dr1, Trihelix) based on published multiple alignments. A total of 56 domains were employed.</p><p>Proteins bearing those domains are regarded as putative transcription factors. Proteins were clustered into 52 families, based on their domain architecture, mainly following Riechmann (2002), and one orphan family. A set of rules were devised in the form:</p><p>Family X = Domain A AND Domain B AND (Domain D OR NOT Domain D) AND NOT Domain C</p><p>The boolean operator "AND" specifies a required relation. Both domains (A and B) must be present in a protein to be assigned to Family X. The operator "AND (... OR NOT ...)" specifies an optional relationship, the domain (D) is not required, though it could appear in some members of the family. The operator "AND NOT" specifies that a domain (C) should not be present in any member of the family. Following this procedure the families obtained are completely exclusive. </p><p>Putative transcription factors classified into families with their accessory domains are available at: <http://ricetfdb.bio.uni-potsdam.de>, <http://arabtfdb.bio.uni-potsdam.de>, <http://chlamytfdb.bio.uni-potsdam.de/> and <http://ostreotfdb.bio.uni-potsdam.de/>. In most cases, upstream, gene, coding and protein sequences are also available. Multiple alignments are available at the domain level for each family. </p><p>For 32 families the presence of a single distinctive domain is the only requirement to assign membership. The

remaining 20 families have different combinations (presence/absence) of 32 different domains. A graph was created to visualize the rules that lead to the definition of each family. This is a bipartite graph, in which one set of vertices represents the transcription factor families, and the other set represents protein domains. There are two types of edges, one type represents a REQUIRED relationship, the other type represents a FORBIDDEN relationship. From this graph the complete set of rules can be easily reconstructed (graph available at: <http://ricetfdb.bio.uni-potsdam.de/v2.0/rules.php>).

In total, 2856 protein models (53 families) harbor domains involved in transcriptional regulation, and are regarded as putative transcription factors in *O. sativa*, 2462 (53 families) in *A. thaliana*, 229 (33 families) in *C. reinhardtii* and 193 (29 families) in *O. tauri*.

We assume that the number of proteins found is a lower bound for the total number of transcription factors in the studied species.

RiceTFDB has been publicly available since June 5, 2004, and it has served more than 6000 different hosts, from around the world.

As of 1st March 2006, a new version (v2.0) of the RiceTFDB is publicly available. From 1st of April ChlamyTFDB, ArabTFDB and OstreoTFDB will be as well.

References

- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32:D138-41.
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14:755-763.
- Riechmann, J.L. (2002). Transcriptional Regulation: A Genomic Overview. pp:1-46. In: Somerville, C. & Meyerowitz, E. (Eds.) *The Arabidopsis Book* American Society of Plant Biologists, Rockville, M