PIBIC Project

Identification of transcription factor genes in plant species: Update of the PlnTFDB (http://plntfdb.bioetanol.cnpem.br/)

Responsible PI: **Dr. rer. nat. Diego Mauricio Riano-Pachon**
CNPEM unit: **Laboratorio Nacional de Ciencia e Tecnologia do Bioetanol (CTBE)**

Background

Organisms are complex systems that interact wit their environment. As any systems they are composed of parts. An important part are the genes making the genome of an organisms. All the cell in a multicellular organism have the same genes, i.e., the same genetic information. However, any organism has different cell types, forming different tissues and organs, i.e., with different functions. This variety of functions can be achieved by the differential control of gene expression, i.e., which genes are turn on or off at any time, thus they have specific spatiotemporal patterns. These patterns are to a large extent established by controlling the transcription of the genes through which RNA copies are generated from the DNA template. In this process, a protein complex composed of general transcription factors (TFs) is mandatory to sustain the expression of all genes encoded by the genome. In addition, other regulatory proteins enhance or repress the transcriptional rate of target genes in response to biotic and abiotic stimuli, and intrinsic developmental processes. These proteins are TFs that bind, in a sequence-specific manner, to *cis*-elements in the target promoters, and other transcriptional regulators (TRs) that exert their regulatory function through protein–protein interactions or chromatin remodeling. The identification of such TFs and TRs from an appreciable number of organisms of divergent lineages represents an important first step towards the understanding of gene regulatory networks and their evolution. We have developed the Plant Transcription Factor Database that is dedicated to the presentation of TFs and TRs and accompanying information of relevance to the research community. The current version of PlnTFDB (v3.0) includes information for 20 different species, it was last updated in June 2010. Since then, tens of new species have had their genome sequence revealed, and they must be included now in PlnTFDB. In addition, advances in computational biology allow us to provide real-time TF and TR identification through PlnTFDB, so this new feature must be added to the current web infrastructure. Thus here we propose to update the identification of TFs and TRs in all the plant (Archaeplastida) genomes that are publicly available, including also species of interest for which only EST data is available.

Objectives

General

To update the Plant Transcription Factor Database (http://plntfdb.bioetanol.cnpem.br/)

Specific

1. To familiarize with the current version of the PlnTFDB and the computational infrastructure to handle the update of the database (Perl scripts and MySQL).
2. To update the set of rules used for the identification of TFs and TRs based on literature revision.
3. To identify TFs and TRs in all the plant (Archaeplastida) species whose genomes are currently

available.

4. To update the MySQL database wit the new identification of TFs and TRs.
5. To implement the function for online classification of unknown proteins in the web server (PHP).
6. To extent the current pipeline in order to apply it to species that only have EST data available.

Methods

Currently the system for the identification of TFs and TRs has more than 150 rules. This systems will be updated and extended through a literature revision. The identification and classification of TFs and TRs is based on a rule system that exploits the domain architecture of proteins. We have seen that some protein domains or domain combinations are exclusive for different protein families. Exploiting this kind of information gives a lot of power to the identification of protein family member, even more when one used probabilistic models of those domain variations through evolutionary time, e.g., hidden markov models (HMM). The elaboration of the rules is a manual process that is based on careful literature curation. Once the rule is found, the next step is to find a HMM that represents the domains in the rule, is not is found in the public database, a new one has to be build and set appropriate threshold for the identification of true hits.

Once tall the rules have associete protein domain models, the rule system can be applied on the deduced proteome of a species of interest, obtaining in that way the identification and classification of TFs and TRs. For more detail please see [1-3].

The computationally deduced proteins for each genome will be downloaded from the appropriate genome resource, e.g., Phytozome, TAIR. PFAM [4] and in-house protein domains will be identified using HMMER [5].

The updated identification of TFs and TRs will be used to populate a MySQL DB that servers as the back-end of PlnTFDB using perl scripts.

Relevant Literature

[1] Riano-Pachón, D. M.; Ruzicic, S.; Dreyer, I. & Mueller-Roeber, B. PlnTFDB: An integrative plant transcription factor database. BMC Bioinformatics, 2007, 8, 42
[2] Pérez-Rodríguez, P.; Riano-Pachón, D. M.; Correa, L. G. G.; Rensing, S. A.; Kersten, B. & Mueller-Roeber, B. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res, 2010, 38, D822-D827
[3] Riano-Pachón, D. M.; Corrêa, L. G. G.; Trejos-Espinosa, R. & Mueller-Roeber, B. Green transcription factors: a chlamydomonas overview. Genetics, 2008, 179, 31-39
[4] Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A. & Finn, R. D. The Pfam protein families database. Nucleic Acids Res, 2012, 40, D290-D301
[5] Eddy S. R. Accelerated profile HMM searches. 2011, PLoS Comp. Biol., 7:e1002195.