Study of Carbohydrate-Active Enzymes (CAZymes), and Transcription Associated Proteins (TAPs) in thousands of genomes across the tree of life

Responsible PI: Dr. rer. nat. Diego Mauricio Riaño-Pachón
CNPEM unit: Laboratório Nacional de Ciência e Tecnologia do Bioetanol (CTBE)

Background

Plants are wonderful harvesters of solar energy, turning it into biomass, e.g., leaves, flowers, stems, roots, etc. That biomass is made of complex polysaccharides, i.e., cellulose, non-cellulosic polysaccharides and lignin. Biomass can be converted into biofuels, such as bioethanol, but in order for this to happen, such complex polysaccharides must be reduced to simpler oligosaccharides and monomers than can be exploited by microorganisms to turn them into bio-products of higher added value. The Carbohydrate-active enzymes (CAZymes) are involved in the assembly and breakdown of these complex carbohydrates and are divided in 286 different families from the classes Glycoside Hydrolases, Glycosyl Transferases, Polysaccharide Lyases, Carbohydrate Esterases, and Auxiliary Activities and additional 74 families of Carbohydrate-binding modules. The expression of these CAZyme-coding genes in a cell is controlled by groups of proteins usually referred to as Transcription Associated Proteins, which include Transcription Factors. There is still a lack of knowledge about which TAPs are involved in the regulation of which CAZyme genes. It has been observed many times that proteins of different families that participate in e.g., the same pathway tend to evolve in a correlated manner, for example have complimentary gains a losses of family members. In this project we plan to explore the complements of CAZymes and TAP families in tens of thousands of genome sequences publicly available, representing the whole tree of life. Exploiting this information we expect to find co-evolutionary dependencies among pairs of families, as well as to identify putative TAPs controlling the expression of CAZymes.

Objectives

- Identify the complete set of cazymes in an automated fashion from complete genome sequences available at GenBank and/or JGI Genomes
- Identify the complete set of transcription associated proteins in an automated fashion from complete genome sequences available at GenBank and/or JGI Genomes
- Employ statistical analysis and comparative phylogenetic methods to evaluate co-evolution among CAZymes and between CAZymes and TAPs.
- Identify expansion and contraction of CAZyme families across the tree of life
- Make detailed phylogenetic analysis of selected CAZyme and TAP families
Methods

CAZymes will be identified using Hidden Markov Models (HMMs) available from dbCAN (http://csbl.bmb.uga.edu/dbCAN/), and TAPs will also be identified using HMMs and a set of rules used to create the PlnTFDB (http://plntfdb.bioetanol.cnpem.br/) and recently extended to include Stramenopiles and Fungi. The rules for the identification of TAPs must be further extended by an extensive review of the literature to include families from Animals, Bacteria and Archaea. This data will be represented in a NxQ matrix, where the N dimension represent the species included in the analysis and the Q dimension represent the families. This matrix will be analyzed in the R statistical package, to generate basic descriptive statistics and evaluate correlation among all family pairs.

During this project the selected candidate will develop abilities for the identification of gene families in a high-throughput manner employing current bioinformatics methods, they will learn how to use High Performance Compute Clusters, they will become familiar with the handling of vary large datasets and comparative phylogenetic methods.

Requirements

The candidate must be an undergraduate student of biological science, computer science, or engineering, they should have basic familiarity of the Linux operative system and have basic knowledge on statistics. Knowledge of phylogenetics methods would be a plus.

Recommended Literature

- Busk PK, Lange L. Classification of fungal and bacterial lytic polysaccharide monooxygenases. BMC Genomics. 2015 May 9;16:368.