Chitinases catalyze the hydrolysis of chitin, a linear homopolymer of β-(1,4)-linked N-acetylglucosamine, which is the second most abundant biopolymer in nature. Based on amino acid sequence similarities these enzymes are classified into glycoside hydrolases (GH) family 18, 19 and 20. GH family-18 and 20 are thought to have a common evolutionary ancestry, since they possess significant similarity in their tertiary structure, catalytic residues and mechanism. GH family-18 is diverse in evolutionary terms and comprises chitinases from bacteria, fungi, viruses, animals and plants. GH family-19 consists of plant, bacteria and some *Streptomyces* chitinases. GH family-20 includes the β-N-acetylhexosaminidases from bacteria, fungi, *Streptomyces* and humans. These chitinolytic enzymes have many industrial and agricultural applications, making their identification and study very promising. This work was a preliminary study that aimed to test a profile Hidden Markov Model (HMM) based strategy in identifying chitinases sequences in metagenomic databases and also to analyze the phylogenetic relationships of the retrieved sequences, the presence of chitinase signatures and the conserved domains variability. In order to perform the analysis, characterized chitinase sequences were searched at UniProtKB database and then fungal and bacterial chitinases of GH family-18, 19 and 20 were retrieved. Since GH family-18 chitinases are highly diverse we split the bacterial and fungal datasets into 9 and 3 subgroups, respectively, according to both the described chitinase subfamilies and a Neighbor-Joining (NJ) tree topology. These characterized chitinase sequences were used to generate 15 datasets in multi-fasta format (12 GH family-18, 1 GH family-19 and 2 GH family-20) which were used as input to run an in-house ruby script pipeline. Briefly, multiple alignments were generated to each chitinase dataset using MAFFT, HMM profiles were constructed using `hmmbuild` from HMMER 3.0 package and `hmmssearch` performed against two environmental metagenomic databases (Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis - CAMERA and Integrated Microbial Genomes – IMG). The best hits (the ones showing the lowest E-value) from each metagenomic project database were parsed out and their fasta sequences retrieved using `fastacmd` from BLAST package. NJ trees were generated for each chitinase dataset with MEGA 5, using the catalytic domain region of both characterized sequences and retrieved metagenomic sequences. The performance of our strategy to identify putative chitinase sequences was evaluated by determining the occurrence of chitinase functional domains, repeats and important sites in the identified CAMERA AND IMG metagenomic sequences using both an InterProScan and a RPS-BLAST search against various conserved domain databases, considering an E-value cut-off of $10^{-5}$. The phylogenetic analysis generated
NJ trees corresponding to each chitinase dataset, making a total of 15 NJ trees. All datasets showed some variability in the amino acid sequence of the catalytic domain region, except for the two active site residues (aspartate and glutamate in GH family-18 and 20, and two glutamates in the case of GH family-19), which were conserved in almost all sequences examined. In addition, the NJ tree analysis also revealed two common sequence patterns, that is, all the 15 trees presented metagenomic sequences phylogenetically related to characterized chitinases - which may help to understand their origin and classification, and all these trees also displayed metagenomic sequences which did not cluster with any characterized chitinase - suggesting a great reservoir of putative new chitinases to be exploited in the metagenomic databases. The \textit{hmmsearch} analysis retrieved a total of 708, 104 and 256 metagenomic sequences from GH family-18, 19 and 20, respectively. The scanning of these sequences using a RPS-BLAST search revealed the presence of chitinase conserved domains in 75\%, 97\% and 98\% of the sequences belonging to GH family-18, 19 and 20, respectively. Similarly, the InterProScan determined the occurrence of chitinase signatures in 82\%, 89\% and 99\% of the metagenomic sequences belonging to GH family-18, 19 and 20, respectively. These results confirmed the efficacy of our HMM-based approach in detecting chitinases sequences. The identification of chitinase conserved domains/signatures also highlighted the great difference in diversity among the three chitinase GH families; while GH family-19 and 20 presented no more than 12 types of conserved domains/signatures, the GH family-18 showed up to 34 types. Interestingly, the small percentage (0.8\% up to 20\%) of the metagenomic sequences showing no hits to any chitinase conserved domain/signature may represent new chitinases that possibly would not be identified using similarity searches. Furthermore, some IMG metagenomic sequences classified as hypothetical protein had hits with chitinase conserved domains in our analysis, indicating that our strategy may have a better performance on sequence annotation when compared to methods based on non-profile similarity searches. The screening tested in this work will be used to perform a more extensive search for chitinolytic enzymes on metagenomic databases by including also sequences of all the other chitinase-producing organisms and it will possibly be used to develop an open access chitinase sequence database, in order to provide functional annotation and information for future studies on these enzymes.

Supported by: CAPES, CNPq and FIOCRUZ