GHEZAIEL Morad – Master 2 Bionformatique, Analyse des données

Genome wide functional annotation of Nanoarcheum Equitans using the EggNOG database

Abstract

Phylogenetic classification of organisms is a key step in comparative genomics. Nowadays, current methods rely on sequence homology inference and functional annotation comparaison. However, querying general databases is time costing when working with complete genomes, and can lead to misinterpretation depending on the alignment methods. In that way, COG databases (Cluster of orhologous groups) come with consistency-based heuristics and have been described as relevant tools for phylogenetic relationship assessments, leading to the development of related databases that takes account of various reigns. In this study, we will take advantage of the EggNOG database to retrieve functional annotation from a thermophilic archae : Nanoarcheum equitans. Particularly, we will assess the efficiency of the EggNOG mapper tool in retrieving functional annotation associated with N.equitans features as thermophilic adaptation and parasitic behavior.

Material and methods :

Complete proteome of N.equitans was pulled from the NCBI server. For EggNOG database querying, we will compare 2 searching methods provided by the EggNOG mapper tool : The HMM (Hidden Markov Models) allows functional annotation of sequences queries by mapping them on clade specific orthologous groups, for N.equitans, we will use the thermococci clade. The second method names DIAMOND allows fast sequence mapping on the EggNOG protein database. Both methods are provided with two stringency criteria : The first allow sequence classification based on multiple alignment with the whole set of orthologous results (coverage) and the second prioritize the classification precision by performing one to one alignment. In order to assess COG based annotation relevance, we will compare these resulting by perfoming a BLASTp on NR databases. For this step, Thermococcales order and Thermococci clade were used as reference NCBI databases. Finally, all the abovementioned results were parsed using a personal python script. Venn diagrams and tables were drawn using respectively the Venn-matplotlib and the pandas library.

Results:

Yield :

Querying the EggNOG database appeared to be relevant for functional annotation. Among the results, 388 and 400 results were respectively obtained using the HMM and the DIAMOND methods. N.equitans proteome being constituted of 536 genes, we retrieved almost 75% of its proteome, making the EggNOG database a relevant tool for genome wide functional annotation. Unexpectedly, both stringency methods (coverage and one-to-one) displayed similar results in term of size and contents. This observation can be explained by the fact that N.equitans is alone in its genus. In addition, Venn diagrams of EggNOG and NCBI NR results (Figure 1) display a large intersection of the results (n=233) but still less than the number of results retrieved using EggNOG, making this database relevant for protein identification.

Content :

N.equitans was described as one of the smallest prokaryotic organism with a genome constituted of half a million base pairs. Previous studies have described this organism as a parasite of another archae named Ignococcus hospitalis. In an evolutionary scenario, N.equitans may have evolved to survive in hydrothermal vents, an environment that can be considered as extreme. In that way, the evolution of N.equitans toward a parasitic life may have been followed by an important genome reduction. Results from functional annotation (Figure 2) allowed us to retrieve important features of this organism. Among them, several amino-acyl tRNA synthetases with proteins associated to tRNA formation. This first observation is consistent with previous studies describing split-tRNA as evolutionary outcomes. In a second hand, we found several proteins associated in DNA reparation (DSBR) and regulation of DNA replication, this result being consistent with the extrem environment where N.equitans lives. As expected, we did not find any functional annotation associated with macromolecule biosynthesis. Conversely, we found annotation that have been described as associated with parasitic life such as glutamate deshydrogenase for energetic metabolism, or thermophilic life such as thermosome formation. Interestingly, results were highly enriched with regulation of transcription, this information could be supported by the fact that prokaryotic genes are highly conserved.

Discussion:

Results from this study are proofs of the relevance of COG based databases. N.equitans is alone in its taxa and studies talks about an evolutionary event that have lead to the apparition of the organism. This hypothesis is strongely supported by the number of results we retrieved using the EggNOG database. In addition, we retrieved the major features of Nanoarcheum equitans such as evidences of its genome reduction, proteins associated with adaptation to its environment and parasitic life. Our study being based on a stringent setup, it’s worthy to say that COG based functional annotation are key tools for comparative genomics. Since the developpement of the first COG database in 2000, several related database has grown up allowing functional annotation of both prokaryotic and eukaryotic genomes. To put it in a nutshell, COG based databases seem to be important genomic tools for large scale analysis compared to conventional blasts.