Protozoa is the common name given to unicellular Eukaryotes organisms that show an extremely diversity. After removing protein redundance with CD-HIT from Plasmodium, Entamoeba, Trypanosoma, Leishmania, Giardia, Theileria, Toxoplasma, Trichomonas and Cryptosporidium species we got 204,624 proteins from 22 proteomes, then submitted to OrthoMCL and 26,101 homologs groups were identified (4,982 paralogs and 21,119 orthologs). 348 groups of orthologous proteins are shared by all the 22 organisms, representing the Protozoa Core Proteome. In this study we discuss the distribution of these homologs genes (paralogs and orthologs) among the protozoa, his function and the functional categorization.
The Protozoa are defined as single-celled eukaryotic organisms showing an extremely diversity and variety. There are more than 200,000 named species of protozoa of which nearly 10,000 are parasitic in invertebrates and almost in every species of vertebrate. The pathogenic species may cause some major diseases (also called Neglected Diseases) such as: malaria, sleeping sickness, Chagas disease, leishmaniasis, amoebiasis and giardiasis, in tropical countries. Comparative studies among Protozoa are important because they may identify similarities and differences leading to the inference of orthologs, paralogs and orphans genes, then we can infer what genes are shared and the ones that are specific to each organism, enhancing our understanding of the biology of each species. The 346,468 sequences from 22 protozoan species used in this study were obtained from NCBI (GenBank/RefSeq) and ProtozoaDB. After removing protein redundancy with CD-HIT from genera Plasmodium, Entamoeba, Trypanosoma, Leishmania, Giardia, Theileria, Toxoplasma, Trichomonas and Cryptosporidium, a total of 204,624 proteins were obtained, then submitted to the OrthoMCL program for homology inference. The number of orthologs clusters, paralogs, orphans genes for each species were identified with the aid of a set of Perl, Ruby and UNIX scripts, as well as some spreadsheet's mathematical functions and the R package. The OrthoMCL generated 26,101 homologs groups (4,982 paralogs and 21,119 orthologs). From such amount, 348 (1.65% groups of orthologous proteins) are shared by all the 22 organisms analyzed, totaling 9,623 entries. Those proteins, represents the Protozoa Core Proteome (PCP) and most of them (90 orthologs representing 25.87%) belong to the functional category known as “J” (Translation, ribosomal structure and biogenesis), followed by the “O” category (Posttranslational modification, protein turnover, chaperones) with 61 orthologous groups (15.53%), the “A” category (RNA processing and modification) with 51 groups (14.65%). Similar analyses, using the functional orthologs group of prokaryotes (COG/NCBI) were done, the “J” category was the most abundant with 106 groups (30.46%), followed by the “R” category (General function prediction only) with 58 orthologous groups (16.67%), the “L” category (Replication, recombination and repair) is the third most common category with 55 groups (15.8%). We also conducted analysis based on score and e-value of their best Blast results of the 348 shared orthologous in order to find out if they were more related to eukaryotes or prokaryotes. Our results shows that 302 groups (86.78%) were more similar to Eukaryotes, 43 groups (12.36%) showed more similarity to Prokaryotes and 3 (0.86%) were equally related to eukaryotes and prokaryotes. Among the PCP we could mention: histone H3, histone H2A, histone deacetylase, actin, alpha tubulin, beta tubulin, 60S ribosomal protein L3/L7/L12/L13a/L13/L17/L18/L32, 40S ribosomal protein S2/S3a/S6/S7/S8/S12/S16, glucose-6-phosphate isomerase, protein disulfide isomerase, valyl tRNA synthetase and asparaginyl-tRNA synthetase. Most of these genes are involved in translation or ribosomal structure. The Kinetoplastida Core Proteome (KCP) has 5,000 orthologous groups, from such amount, 3,396 (67.92%) are exclusive to them (Kinetoplastida Specific Genes - KSG), then do not have similarity with genes in others taxa, although genes with similar functions can be found. KSG contains a large amount (1,592/3,396 or 46.29%) of genes annotated as hypothetical,and from the ones with function determined, we can mentions 20 genes: aminopeptidase, ADP-ribosylation factor, hexose transporter, lipase, kinesin, calpain-like cysteine peptidase, fatty acid desaturase, Zn-finger protein, 40S ribosomal protein L14, 30S ribosomal protein S17, 50S ribosomal protein L13, 60S ribosomal protein L28, glutathione peroxidase, actin-like protein, p21 antigen protein, cytochrome p450 reductase, ABC transporter, dynein arm light chain, lathosterol oxidase and chaperone protein DNAJ. A functional characterization using KOG/NCBI was made using the KSG, then observed that the functional category “R” is the most abundant with 456 groups (13.43%). The second most abundant category was "L" with 198 groups (5.83%), followed by the "T" category (Signal transduction mechanisms)
with 188 groups (5.54%). The Apicomplexa Core Proteome (ACP) was formed by 986 orthologous groups, where 224 (27.82%) are Apicomplexa Specific Genes (ASG), in other words, they have a function that is similar to other organisms, but have no similarity to these genes found in other groups, have similarity only among the Apicomplexa, and so are classified as ASG. Some of them are: helicase, histone binding protein, heat shock protein 70 (hsp70), heat shock protein 90, RNA helicase, RNA methyltransferase, zinc finger protein, and many hypothetical proteins. The most abundant functional category in ASG was “R” with 55 groups (24.55%), followed by category “K” (Transcription) with 20 groups (8.93%) and “O” with 18 groups (8.04%). The Protozoa species that presented the highest number of protein paralogs and paralogs groups, was Trichomonas vaginalis, with 35,650 proteins and 2933 groups. It presented an average of 12.15 protein per group (35,650/2,933), and the functional classification shows that the most abundant category was the "T" category with 469 groups, followed by the "O" category with 281 groups, then "M" (Cell wall / membrane / envelope biogenesis) with 83 groups. The organism that presented the highest gene family expansion was the P. knowlesi with 298 proteins from 16 paralogs groups, having an average of 18.63 protein per group (298/16). Using KOG functional categorization, the "T" category is the most abundant with 3 groups, followed by "Z" (Cytoskeleton) with two groups. Trypanosoma cruzi had 814 paralogs groups containing 5,088 proteins in total and an average of 6.25 proteins per group, the most abundant category for this paralogs was "T" with 60 groups and the second category was "M" with 12 groups, followed by the "V" category (defense mechanisms) with 10 groups. The organisms that have the smallest number of paralogs proteins and paralogs groups were those of the genus Cryptosporidium, C. hominis and C. muris. They showed four paralogs proteins in two groups, while C. parvum showed seven proteins distributed in three groups. For the rest of the groups, there are no patterns in number of paralogs proteins and paralogs groups within the genus, for example, the genus Plasmodium showed values of protein per group (average of the paralogs genes) ranged from 18.63 in P. knowlesi, to 2.34 in P. berghei. We can observe that the PCP genes are generally related to the maintenance of the cell and information processing, because were found ribosomal proteins, histones, cytoskeletal proteins and tRNA synthetase. Moreover most of the genes are distributed in functional categories "J", "A" and "O", confirming the idea that these genes are more related to maintenance and processing, falling into two divisions: "Information Storage and Processing" and "Cellular Processes and Signaling". Among KSG, it was noted genes with more specific functions, such as ABC transporter related to drug resistance; carbohydrate metabolism: hexose transporter; corroborating that Kinetoplastida have a different way of carbohydrate metabolism. Among the ASP genes, it was observed they have a particular way of DNA/RNA processing because we found genes that are related to them, such as: RNA helicase, RNA methyltransferase and histone binding proteins. However most of the KSG (46.29%) and ASP (40.63%) have hypothetical function, which is somewhat expected, because these are specific genes. Because the nature of KSG, it is not possible to transfer annotation by similarity, since there are no similar genes in other previously studied organisms. Regarding the paralogs genes observed for T. vaginalis, the most abundant categories were "O" and "M". As hypothesized in the literature these many copies help the organisms to produce a larger amount of protein and a more diverse repertoire of surface proteins aiding evasion of the immune system. In T.cruzi, it is interesting to note that "M" and "T" categories are present, and it is known that this species have a large family of surface proteins (Mucins) and moreover, they are described to be related to defense and invasion mechanisms, then such a large amount of copies is expected because it increases the repertoire of defense, enabling the exchange of these proteins, our findings corroborate these observations.