Inference of distant homologs in Protozoa by pHMM–pHMM comparison for the identification of superfamilies

Darueck Campos, Rodrigo Jardim, Alberto M. R. Dávila

Oswaldo Cruz Institute, Acre Federal Institute

According to World Health Organization, the major diseases in tropical countries, such as malaria, sleeping sickness, Chagas disease, leishmaniasis, amebiasis and giardiasis, are causing by protozoan parasites, which together threaten more than a quarter of the world population. In recent years, as a result of the work of several research teams, 71 Protozoa species were fully sequenced, but a majority portion of their proteins have not been functionally annotated yet. The use of pHMM (profile Hidden Markov Model) for identifying distant orthologs in those Protozoa is considered more efficient than other techniques such as comparison of protein sequences or between pHMM and protein sequences. The main reason is its potential to discover more distant homologs. Furthermore, this methodology may also contribute to the functional annotation of proteins thus enabling the improvement of knowledge about the species under study. In theory, distant homologues identification might result in the protein superfamilies identification. In light of this, we aimed to identify superfamilies by analyzing 3 Protozoan genomes: *Cryptosporidium muris*, *Entamoeba invadens* and *Trypanosoma grayi*, chosen for their evolutionary distance, using pHMM- pHMM. This methodology, was able to identify 94% of distant orthologs among all the orthologous groups inferred from the three species. Considering only 2 species, our methodology was able to identify an average of 75% of distant orthologs between *C. muris* and *T. grayi*, 50% between *C. muris* and *E. invadens* and for *T. grayi* and *E. invadens* we found 60% of distant orthologs. Our results are encouraging and allow the annotation of proteins based on distant homology inference.