POTION: a massive parallel program for identification of homologous genes under positive selection on genomic scale datasets

Jorge Hongo, Giovanni Castro, Felipe Silva, Leandro Cintra, Adhemar Zerlotini, Francisco Lobo

UNICAMP, Instituto de Computação, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária

A considerable amount of the genes do not have a known biological function, and constitutes a rich field for the identification of genes of interest for biotechnological applications, ranging from the discovery of new drug targets to the search of genes responsible for phenotypes of agricultural interest. The majority of amino acid residues in multiple alignments of homologous proteins do no vary as expected by mutational rates alone, but rather, tend to be highly conserved. When analyzing the respective homologous codons of such alignments, a much higher-than-expected synonymous mutation ratio is observed. This phenomenon occurs since a non-synonymous mutations usually decrease the fitness of the organism, and these alleles are removed from populations by purifying selection. However, some groups of homologous genes present a higher-than-expected ratio of non-synonymous mutations or, in other words, a selective pressure for variation instead of conservation, a phenomenon known as positive selection. Genes under evidence of positive selection are of particular interest due to their close association with important elements of the ecological niche of the specie under analysis, such as the parasite genes involved in host-parasite biological interaction. Therefore, the search for positive selection could help prioritize unknown genes for further characterization. However, several of the computational steps needed for positive selection detection are time-consuming, such as codon model likelihood calculation and phylogenetic tree reconstruction. The search for positive selection is also highly parallelizable, since each group of homologous can be analyzed independently. Here we describe POTION (POsitive selecTION), a massively parallel software to automatically detect positive selection on genomic scale data, ideal for analysis of large datasets in cluster infrastructures. POTION is highly modular and easily expandable, and currently uses state-of-the-art software for protein sequence alignment (MUSCLE and prank), sequence trimming (trimAl), phylogenetic tree reconstruction (phylip) and codon model evolution likelihood calculation (PAML). To validate POTION, we used as gold-standard a dataset of 40 groups of paralogs from Trypanosoma brucei previously surveyed for positive selection. This study found 23 groups of paralogs to be under positive selection. POTION could detect positive selection in 22 out of the 23 known positively selected genes, misclassifying only one non selected gene as positively selected, with precision, recall and F-measure values of 0.96. This analysis also followed a linear time decrease proportional to the number of processors allocated until capped by the longest computational step for an individual group. The high F-measure observed, together with the linear decrease of the computational time, demonstrates that POTION could effectively be used in a wide range of bioinformatics studies. POTION is a first-of-its-kind software that will potentially hasten projects where genome-scale positive selection identification is a key factor. Supported by: CNPq

Keywords: comparative genomics, positive selection, parallelization, computer cluster, genomic-scale analysis

Concentration area: Genomics Evolution

⬅