POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes

Jorge Hongo, Giovanni de Castro, Leandro Cintra, Adhemar Zerlotini, Francisco Lobo

Embrapa Informática Agropecuária

Abstract

Detection of genes evolving under positive Darwinian evolution in genome-scale data is nowadays a prevailing strategy in comparative genomics studies to identify genes potentially involved in adaptation processes. Despite the large number of studies aiming to detect and contextualize such gene sets, there is virtually no software available to perform this task in a general, automatic, large-scale and reliable manner. This certainly occurs due to the computational challenges involved in this task, such as the appropriate modeling of data under analysis, the computation time to perform several of the required steps when dealing with genome-scale data and the highly error-prone nature of the sequence and alignment data structures needed for genome-wide positive selection detection. We present POTION, an open source, modular and end-to-end software for genome-scale detection of positive Darwinian selection in groups of homologous coding sequences. Our software represents a key step towards genome-scale, automated detection of positive selection, from predicted coding sequences and their homology relationships to high-quality groups of positively selected genes. POTION reduces false positives through several sophisticated sequence and group filters based on numeric, phylogenetic, quality and conservation criteria to remove spurious data and through multiple hypothesis corrections, and considerably reduces computation time thanks to a parallelized design. Our software achieved a high classification performance when used to evaluate a curated dataset of Trypanosoma brucei paralogs previously surveyed for positive selection. When used to analyze predicted groups of homologous genes of 19 strains of Mycobacterium tuberculosis as a case study we demonstrated the filters implemented in POTION to remove sources of errors that commonly inflate errors in positive selection detection. A thorough literature review found no other software similar to POTION in terms of customization, scale and automation. To the best of our knowledge, POTION is the first tool to allow users to construct and check hypotheses regarding the occurrence of site-based evidence of positive selection in non-curated, genome-scale data within a feasible time frame and with no human intervention after initial configuration. Our software was recently published in BMC Genomics (http://www.biomedcentral.com/1471-2164/16/567) and is available at `http://www.lmb.cnptia.embrapa.br/share/POTION/`. This work was supported by Embrapa (Brazilian Agricultural Research Corporation), LMB (Laboratório Multiusuário de Bioinformática) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [grant number 485279/2011-8].