An integrated approach to building and applying profile HMMs for sequence detection in genomic and metagenomic data

Liliane Santana Oliveira Kashiwabara, Arthur Gruber

USP

Abstract

In this work, we report the development of an integrated approach for the construction of profile HMMs and their use in a series of applications using genomic and metagenomic data. To construct profile HMMs, we developed TABAJARA, a program for the rational design of profile HMMs using multiple sequence alignment (MSA) files. By optimizing a position-specific information score in a sliding window along the length of an MSA, TABAJARA automatically identifies the most informative sequence motifs that are either (1) conserved across all sequences or (2) discriminative for two or more specific groups of sequences, and then constructs profile HMMs from these alignment blocks. The generated models can be used to screen genomic or metagenomic sequencing data with HMM-Prospector program, a tool that performs similarity searches and quantifies the results according to score or e-value thresholds for each tested profile HMM. Models displaying the most relevant results can be used as seeds by GenSeed-HMM, another tool developed by our group to perform seed-driven progressive assembly using unassembled sequencing data. Finally, profile HMMs can also be used by the program e-Finder to identify and extract multigenic elements, starting from assembled genomes or metagenomes. Potential applications include the detection of proviruses, mobile genetic elements or any other set of specific genes present in a specific syntenic context, such as operons. We obtained a successful set of results using this toolbox for studying casposons, a family of self-synthesizing mobile elements that are found in archaeal and bacterial genomes and that gave rise to the currently known CRISPR elements. First, we designed casposon-specific profile HMMs constructed from endonuclease Cas1 and DNA polymerase B sequences. These models were used to screen several unassembled metagenomic datasets with HMM-Prospector. The positive sets were submitted to progressive assembly with GenSeed-HMM program, using the profile HMMs as seeds, resulting in the reconstruction of casposon-specific sequences. Also, the same models were used with e-Finder to detect and retrieve casposons sequences from assembled bacterial and archaeal genomes of the PATRIC database. A total of 138 elements derived from 105 distinct genomes were detected. This number of elements is approximately three times higher than the number of casposons reported in the literature. In both cases, phylogenetic analyses confirmed the correct taxonomic assignment of the positive sequences.

Funding: CAPES