Automatic identification of horizontal gene transfer events in bacterial genomes

Manolov A.[1], Kanygina A.[2]

[1]Research Institute of Physical and Chemical Medicine, Moscow

[2]Moscow Institute of Physics and Technology, Moscow

Horizontal gene transfer (HGT) is a major force of bacterial evolution. According to research, the percentage of bacterial genes acquired by HGT is approximately 10-15% of the total number of genes. Identification of such genes is extremely important for deeper understanding of the evolutionary processes in bacteria; thus, the task of fast and reliable search for HGT events is seems actual.

Now there is a number of existing algorithms performing identification of HGT in a genome. Some of them are based on the multiple alignment of genetic sequences, other perform the search for HGT de novo (having no reference sequence). These algorithms employ the fact that some of the characteristics of HGT areas such as GC content or codon usage differ from the average of the genome. However, such an approach often needs further verification. In this paper we suggest an algorithm that combines two approaches: we use both statistics and reference sequences.

The algorithm is implemented in the form of a pipeline based on a set of BioPerl scripts. We have employed a number of utilities and Web services, such as BLAST+, BLAST (http://blast.ncbi.nlm.nih.gov/), Circos. As an input the program requires a nucleotide sequence of the query organism (FASTA format), its annotation (GenBank) and a set of reference sequences (closely related organisms or strains). These reference sequences can be identified and downloaded automatically (as an option). The first step is constructing a BLAST database of these sequences. Next, all the coding sequence (CDS) from the query are aligned against this database, and some CDS show no matches with the reference. The next stage is the alignment of such CDS globally against the NCBI database. As a result, the most frequently matched organisms are included in a table sorted by relevance, with the indication which query region they belong to. GC content and codon usage are then calculated for these regions to decide, whether they are probably acquired by HGT. The final step is the visualization of the alignment and putative HGT with Circos utility.