RapH and RapD: two indexes designed for de novo identification of repeats in whole

Campagna D. ,Romualdi C. ,Vitulo N. ,Favero M. ,Lexa M. ,Cannata N. and Valle G.

CRIBI, Università degli Studi di Padova

Motivation

The identification of repeats is an essential step for genome analysis and annotation, but is not easy because repeats tend to be little conserved during evolution. This particular aspect of repeats makes very difficult the identification of homologous sequences that diverged significantly, both within the same genome and between genomes of different organisms.

Methods

We have recently developed RapH, a program specifically designed for de novo identification of repeats in whole genomes. We have implemented this program both for 32 and 64 bits architecture machines. The last one is optimized to run with mammalian genomes 2-3 Gbases long. RapH has an execution time proportional to the length of the genome and is based on an algorithm that in the first step counts all the possible gapped and ungapped words of a given length (typically log4 of the genome length) and in the second step reports on the genome how often any given word was found (i.e. the RapH score). Therefore the general strategy of RapH is based on counting the occurrences of gapped and ungapped words of a given length in a genomic sequence. Furthermore, we have developed a second algorithm called RapD, that is also based on the counting of gapped words. However, instead that counting the occurrences of words, RapD considers the distance relations of contiguous gapped pattern. RapD is extremely sensitive for finding divergent duplicated sequences, especially when they occur close to each other.

Results

The scores obtained by RapH and RapD give a different measure of the "repetitiveness"; therefore it can be useful to display both results at the same time. For this purpose we have developed RapTools, a program to handle and visualize RapH and RapD results together with other information such as the linguistic complexity, chaos game representation, GC content and Dot Plot. Finally, the human genomic sequence has been analyzed by means of RapH and RapD, revealing interesting features. The resulting scores have been integrated in the UCSC Genome Browser to allow the better comparison with other genetic

Contact email: davide@cribi.unipd.it
URL: http://genome.cribi.unipd.it/