Development and usage of a new bioinformatic strategy for identifying repeated sequences in genomic DNA

D. Campagna, G. Valle

CRIBI, Università di Padova

Repeated sequences are quite common in the genomes of most organisms, from bacteria to plant and animals, but they may be quite different from organism to organism both in amount and kind. While most repeated sequences from the best studied organisms have been identified and classified, very few approaches have been described for a systematic detection of repeated sequences in new genomes. The aim of the work described in this presentation was to develop a computer program to perform this specific task.

Our approach is divided into two phases: firstly, the repeated sequences are identified on the genome; secondly, they are analyzed and characterized. During the first phase a wide series (typically several million) of different small patterns are counted on the genomic sequence and a simple statistical analysis is performed to identify all the patterns occurring at an unexpectedly high frequency. All these patterns, one by one, are subsequently re-positioned on the genomic sequence where counters are placed at each nucleotide and are incremented every time a base of a frequent pattern is found to occur at a specific position. As a result the regions corresponding to repeats will be easily recognizable as those with the highest counts.

The second phase of the process is the recovery of all the repeats from the genomic sequence, followed by an assembly procedure that clusters together all the similar repeats and allows their characterization.

The entire strategy and the individual programs have been tested to make an evaluation of the performance and suitability for achieving the initial aim. Several tests have been performed on random sequences, typically 4 million bases long, as well as on the same sequences containing given numbers of repeats of known length. We could establish that the program can easily identify 3 repeats of 100 bases embedded in a 4 million bases long random sequence. The same sensitivity can be obtained using a genomic sequence instead of a random sequence.

A further control was done on 4 million bases of human genome, comparing the results obtained by our program with those of RepeatMasker that is able to detect known repeats given in a database. All the repeats identified by RepeatMasker were also identified by our program that also found a few other repeated sequences missed by RepeatMasker, including some low complexity regions and a few local repeats. The comparison against RepeatMasker has also shown that our program is able to complete a typical analysis of a 4 million bases genome in a few minutes, while RepeatMasker takes several hours to complete the analysis on the same sequence. It mast be pointed out that RepeatMasker requires a list of the repeats to be searched, while our program uses an intrinsic approach that does not require any other information than the sequence to be analyzed.

In conclusion, the velocity of execution, the independence from the complexity of the repeats and the intrinsic approach make our program ideally suited for the systematic identification of repeats in new genomes.