COMPUTER ANALYSIS OF MULTIPLE REPEATS IN BACTERIA

Vitreschak A.*, Noe L., Kucherov G.

INRIA-Lorraine/LORIA, 615, rue du Jardin Botanique, BP 101, 54602 Villers-lès-Nancy, France
* Corresponding author: e-mail: vitresch@loria.fr

Keywords: *repeats, clusterization, bacteria*

Summary

Motivation: The presence of repeated sequences is a well-known feature of bacterial genomes and interpretation and classification of those repeats is an actual problem.

Results: We described a method for computing *multiple repeats*, that is sequences that have multiple (two or more) occurrences in a genome. In order to identify multiple repeats in bacteria genomes, we apply the YASS software (Noe, Kucherov, 2004) and developed a novel algorithm for multiple repeat clusterization. Exhaustive computation and analysis of those “clusters of repeated sequences” in bacteria is the subject of the present work.

Availability: Program is available by e-mail: vitresch@loria.fr

Introduction

The presence of repeated sequences is a well-known feature of bacterial genomes. In general, a DNA repeat is a sequence, which appears at least in two copies in the genome. The size of repeated sequences and their biological function differ greatly: in one case, a repeat can be about a thousand nucleotides long and contain coding open reading frames (for example, a mobile element); in other cases, a repeat can correspond to a regulatory element located in intergenic regions. Moreover, repeated sequences can be strongly conserved not only within one genome, but also across different (in some cases remotely related) genomes.

There are several programs specially devoted to the computation of repeats within a given genomic sequence (Kurtz *et al*., 2001; Vincens *et al*., 1998; Lefebvre *et al*., 2003). Alternatively, such repeats can be obtained by computing, using any local alignment method, local similarities between the input sequence and itself. On the other hand, there is no method to systematically compute *multiple repeats*, that is sequences that have multiple (two or more) occurrences in a genome. Exhaustive computation and analysis of those “clusters of repeated sequences” in bacteria is the subject of the present work.

Data and Methods

In order to identify multiple repeats in a bacterial genome, we first apply the YASS software (Noe, Kucherov, 2004) and find all strong local similarities, viewed as two-copy repeats, within the genome. YASS parameters have are set to detect 70 % similarity alignments with a very low false positive rate (using the seed ##@_#@_#_#__## of weight 8 and group size 11, for details see Noe, Kucherov, 2004). All possible repeated sequences found by YASS are then grouped into clusters, with the goal that each cluster contains all copies of the same repeated biological element.

The clusterization of possible repeats is made in two steps. The first step (pre-clustering) consists in processing all local alignments found by YASS. This pre-clustering step groups together sequences that are strongly related: this is achieved by a heuristical search for quasi-cliques (almost perfect cliques) in the graph in which nodes are sequences and edges are similarities. The data structure used at this step is an interval tree that stores the coordinates of each sequence occurring in each YASS alignment. These initial clusters are “starting points” for further clusterization and are essential for the stability of “cores” of clusters (see below).

A method of "cores" is used at the second step of clusterization. Its main idea consists in using most conserved parts of repeats, called "cores", for controlling the clusterization process. First, a graph is constructed with nodes corresponding to the initial clusters. An edge connects two nodes when at least one sequence from one initial cluster "overlaps" at least one sequence from another initial cluster. Additional conditions for connecting two nodes (initial clusters) are the following:

$$\min (L_1/L_{overlap}, L_2/L_{overlap}) < 2 , \qquad (1)$$

$$\max (L_1/L_2, L_2/L_1) < 2, \qquad (2)$$

where $L_{1,}$ L_1 and $L_{overlap}$ are the lengths of first repeat, second repeat and the length of common part (overlap), respectively.

The first rule means that the length of the common part is at least a half of the minimal length of the two repeats. The second condition insures that the two involved sequences have comparable lengths.

Detected repeats from initial clusters correspond either to an entire repeated element or only to its part. In some cases, repeated elements correspond to a superposition of two or more different adjacent repeats (sometimes partially overlapped). This is an additional difficulty for the appropriate detection of repeated units. For example, if only rules (1), (2) are used for clusterization, then the process can result in a huge cluster containing more than 95 % of initial clusters (as applied to the *Neisseria meningitidis* genome). This fact is due to adjacent locations of distinct repeated elements on the DNA sequence, that can erroneously fall into one cluster. A simple illustration is given in Figure 1. The initial cluster1 is joined with the initial cluster2 and the latter is joined with the initial cluster3. cluster2 contains parts of both repeat1 and repeat2 and because of this "bridge", initial clusters 1 and 2 are also joined. In this way, different non-related repeats can be joined together, and the whole process results in one huge "supercluster". To cope with this problem, a method of "cores" has been developed.

At the second step, a "core interval" (core) is computed for each cluster. The core corresponds to the most conserved part of the repeat and core coordinates are computed as the average of corresponding sequence coordinates of the cluster.

Using the cores, the clusterization step is defined as the following traversal of the set of clusters (Fig. 1A): (a) after constructing the set of initial clusters, choose a start initial cluster (the largest one) (b) iteratively join the current cluster with other clusters which verify rules (1), (2) *applied to cores*. Manipulating cores allows us to avoid joining unrelated clusters, as shown in Figure 1. Figure 2B illustrates that those clusters are not joined anymore since rules (1), (2) are not verified for cores.

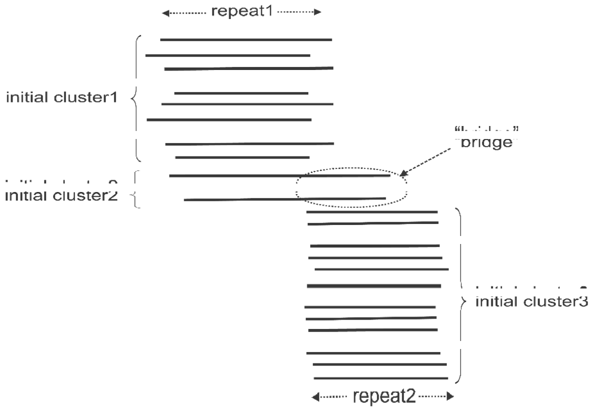

Fig. 1. Using the cores, the clusterization step is defined as the following traversal of the set of clusters (Fig. 1A): (a) after constructing the set of initial clusters, choose a start initial cluster (the largest one) (b) iteratively join the current cluster with other clusters which verify rules (1), (2) *applied to cores*. Manipulating cores allows us to avoid joining unrelated clusters, as shown in Figure 1. Figure 2B illustrates that those clusters are not joined anymore since rules (1), (2) are not verified for cores.

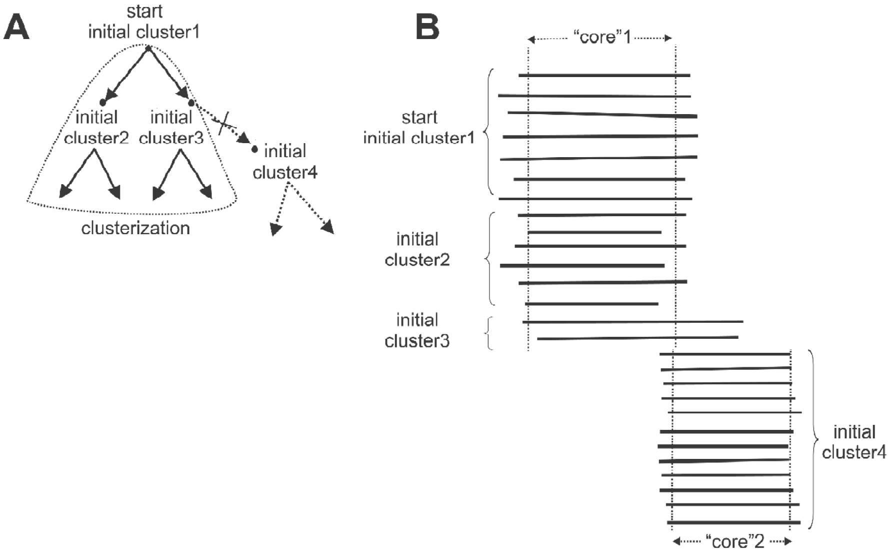

Fig. 2.

Results

We run our method on the *Neisseria meningitidis* genome and obtained a number of interesting clustered repeats, some of them with a known well-identified biological function. Interestingly, one resulting cluster embraced several hundreds of ρ-independent terminators. Several other clusters corresponded to mobile IS-elements (IS30, IS1016C2, IS1106).

Besides of those known elements, some interesting unknown repeats have been detected. For example, we found a cluster of sequences of about 120 bp long, which are highly distributed in the genome (more than 100 copies). These repeated sequence has a complex palindromic structure and is located in intergenic regions only, which suggests its possible regulatory role. Alternatively, this repeated element might be an RNA with a strong secondary structure, or a short mobile element of a new kind (suggested by its high degree of distribution). Another complex repeated element, revealed by our procedure, are also located in non-coding regulatory regions often adjacent to genes involved in bacterial pathogenesis. This demonstrates that the proposed clusterization method allows us to detect new repeats with unknown biological function. Interpretation and classification of those repeats is the subject of our current work. Program is available by e-mail:vitresch@loria.fr.

References

Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale // Nucleic Acids Res. 2001. V. 29. P. 4633–4642.

Lefebvre A., Lecroq T., Dauchel H., Alexandre J. FORRepeats: detects repeats on entire chromosomes and between genomes // J. Bioinformatics. 2003. V. 19. P. 319–326.

Noe L., Kucherov G. YASS: enhancing of sensitivity of DNA similarity search. submitted to BGRS-2004.

Noe L., Kucherov G. YASS: enhancing of sensitivity of DNA similarity search // Research report RR-4852, INRIA. http://www.inria.fr/rrrt/rr-4852.html [In French]. 2004.

Vincens P., Buffat L., Andre C., Chevrolat J.P., Boisvieux J.F., Hazout S. A strategy for finding regions of similarity in complete genome sequences // Bioinformatics. 1998. V. 14. P. 715–725.