

Bacterial genomes are internally formatted

Korotkov E.V.*

Institute of Bioengineering, Research Center of Biotechnology, RAS, Moscow, Russia

* bioinf@yandex.ru

Key words: genome; dispersed repeats; bacteria

Motivation and Aim: The goal was to try to find dispersed repeats in the bacterial genome. It is known that the bacterial genome is packaged into a nucleoid through the interaction of DNA with certain proteins. Due to this packaging, it can be assumed that weakly similar families of dispersed repeats may exist in the bacterial genome, which could not be detected by previously developed computer methods. Therefore, we decided to look for dispersed repeats in bacterial genomes in the range $1.0 \leq x \leq 1.7$, where x is the average number of mutations per nucleotide between any two repeats from a family.

Methods and Algorithms: We have developed a de novo method for the identification of dispersed repeats based on the use of random position-weight matrices (PWMs) and an iterative procedure (IP) [1]. The created algorithm (IP method) allows detection of dispersed repeats for which the average number of substitutions between any two repeats per nucleotide (x) is less than or equal to 1.7. IP method made it possible to detect families of dispersed repeats in bacterial genomes which have not been previously reported. The IP method uses positional weight matrices, dynamic programming and an iterative procedure. Z , which is approximately an argument for a normal distribution, will be used as an estimate of statistical significance.

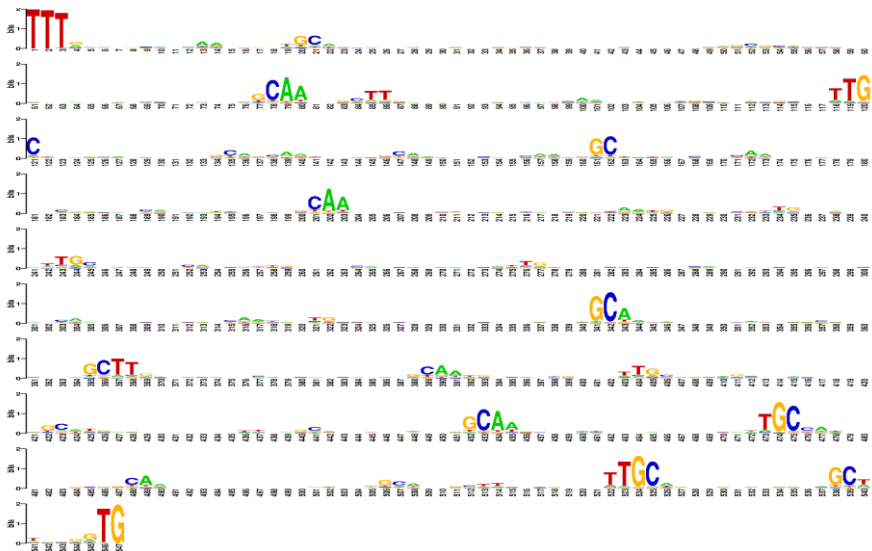


Fig. 1. Consensus sequence for the dispersed repeat family from the *Clorobium chlorochromatii* genome constructed using Weblogo [2]

Results: We applied this method to find dispersed repeats in the genomes of *E. coli* and 42 bacteria from other bacterial phyla and could identify repeat families comprising from 1.0×10^3 to 1.4×10^4 , depending on the species. Each bacteria contains only one family of dispersed repeats with lengths from 440 to 580 bases. The repeats found occupy from 30 to 60 % of the bacterial genome, and more than 90 % of the repeats occur in the coding sequences of the genome. They are superimposed on the coding region as a motif. Such extensive repeat family could not be detected in the analyzed genomes by using the RED, RECON, or Repeat_masker programs but only by the IP method, which could find de novo repeat families with $x \leq 1.7$, whereas all other programs could do it with $x \leq 1.0$.

In Figure 1 we show the consensus sequence that was constructed using the WebLogo program [2] for a family of repeats from the *Clorobium chlorochromatii* genome. The figure shows that dispersed repeats of this family diverged greatly from each other. However, one can notice a fairly large number of short regions of 2-5 bases in length, which are almost identical in all repeats. The existence of such repeats shows that bacterial genomes are marked up like a computer hard drive. This markup could be involved in the creation of the liquid crystal structure within bacterial DNA through interactions between repeats within a family and some proteins [3].

Funding: The study is supported by RNF grant 24-24-00031.

References

1. Korotkov E., Suvorova Y., Kostenko D., Korotkova M. Search for dispersed repeats in bacterial genomes using an iterative procedure. *Int J Mol Sci.* 2023;24:10964. doi 10.3390/IJMS241310964/S1
2. Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. WebLogo: a sequence logo generator. *Genome Res.* 2004;14:1188-1190. doi 10.1101/GR.849004
3. Yevdokimov Y.M., Salyanov V.I., Skuridin S.G. From liquid crystals to DNA nanoconstructions. *Mol Biol.* 2009;43:284-300. doi 10.1134/S0026893309020113/METRICS