Avoided and Under-Represented Strings in Bacterial Complete Genomes — Some Associated Fractal and Combinatorics Problems

Bailin Hao
Institute of Theoretical Physics and
Beijing Genomics Institute
Academia Sinica, China
hao@itp.ac.cn

Abstract

The availability of more and more bacterial complete genomes makes it possible to ask many global questions. One of the simplest such questions is whether there are avoided short strings in one or another genome. We use a simple visualization scheme as well as direct counting to explore the situation. Each bacterium shows a specific pattern in the visualization scheme and some clearly avoid a certain set of palindromic short strings. We discuss possible biological implication of the observation and possible application of the method. The fractal-like patterns in the visualization scheme become well-defined fractals in the infinitely long string, hence non-biological limit. The direct counting leads to a problem as how to distinguish true and redundant avoided strings. The latter is associated with the exact calculation of the fractal dimensions in the above mentioned limit. This problem is solved by using the Goulden-Jackson cluster method in combinatorics and by constructing the minimal deterministic automaton for the factorisable language defined by a complete genome.

References

- B.-L. Hao, H. C. Lee and S.-Y. Zhang, "Fractals related to long DNA sequences and bacterial complete genomes", Chaos, Solitons and Fractals 11 (2000) 825 – 836.
- B.-L. Hao, "Fractals from genomes — exact solutions of a biology-inspired problem", Physica A282 (2000) 225-246.