DOI 10.18699/WIBSB-2018-49

Fast search of long approximate repeats in DNA sequences with bounded indel density

S. Tsarev[1]*, M. Senashova[2], M. Sadovsky[1, 2]
[1] *Siberian Federal University, Krasnoyarsk, Russia*
[2] *Institute of computational modelling SB RAS, Krasnoyarsk, Russia*
** e-mail: sptsarev@mail.ru*

Key words: algorithms, Vernier pattern, edit distance, gauge repeat, mutation

Motivation and Aim: The search of common strings in two or several symbol sequences makes a core in bioinformatics and up-to-date molecular biology. The problem is far from a completion, in spite of a long story. In general, the problem is the following: given sequences T_1, T_2, …, T_k of symbols from some finite alphabet, find all possible common substrings (i. e. coherent subsequences) occurred in the sequences, maybe, with some mismatches. Previously, a new algorithm for the fast search of common substrings in two or several symbol sequences had been reported [1]. The algorithm was originally implemented for the exact matching strings search, while it allows some extensions for error tolerant search of substrings. The algorithm [1] for search of exactly matching substrings is much faster compared to the brute force search methods; it is based on a simple idea of rarefied dictionaries and uses the classical Vernier scale, cf. for example [2].
Results: A novel algorithm to find all sufficiently long repeating nucleotide substrings in one or several DNA sequences is proposed. The algorithm searches approximately matching strings very fast with given level of local mutation density. Also, the extended version of the method to identify all sufficiently long repeating nucleotide substrings in one or several DNA sequences with indel mismatches is proposed. The method based on a specific gauge applied to DNA sequences that guarantees the identification of all repeating substrings. The method allows the matching substrings to contain a given level of errors of all types. The gauge is based on the development of a heavily sparse dictionary of repeats, thus drastically accelerating the search procedure [1–3]. Some biological applications illustrate the method.
Acknowledgements: This study was supported by a research grant No. 14.Y26.31.0004 from the Government of the Russian Federation (M.G. Sadovsky) and the grant from Russian Ministry of Education and Science to Siberian Federal University, contract No. 1.1462.2014/K (S.P. Tsarev).

References

1. Tsarev S., Sadovsky M. (2016) New error tolerant method for search of long repeats in DNA sequences. LNBI. 9702:171-182.
2. https://en.wikipedia.org/wiki/Vernier_scale
3. Tsarev S., Senashova M., Sadovsky M. (2018) Fast Algorithm for Vernier Search of Long Repeats in DNA Sequences with Bounded Error Density. LNBI. 10849:88-99.