Comparison of the Alignment and Shaidurov's Methods the Search Efficiency in Symbol Sequences with Mismatches

Anna Molyavko[1], Evgenia Karepova[2], Mikhail Sadovsky[2,4], Igor Borovikov[3] and Olga Mutovina[1]

[1]*Siberian federal university, Krasnoyarsk, Svobodny pr., 79, 660041, Russia*
[2]*Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia*
[3]*Nekkar.net LLC, Foster City, CA 94404*
[4]*Federal Research & Clinic Center of FMBA of Russia, Kolomenskaya str., 26, Krasnoyarsk, 660037, Russia*

Abstract

The new method for comparison and analysis of symbol sequences is proposed; the method is based on the convolution function calculation defined over the binary numeric sequences derived from the original symbol sequence. The method provides highly parallel implementation and is very powerful in insertion/deletion mutations search. A discrete fast Fourier transform is implemented for convolution calculation. Also, an idea of the alphabet expansion is proposed to improve the signal/noise ratio. Some genomic applications are provided and discussed. The applications are used to illustrate and overcome the problem of signal/noise selection, and alignment localization.

Keywords

Pattern recognition, anomaly detection, parallel computation, InDel-genome comparison, knowledge retrieval

1. Introduction

Comparison of symbol sequences stands behind the methodology in many fields of science, ranging from mathematics to bioinformatics and linguistics. There are two versions of the problem: the former is an exact matching search, and the latter is a homology search with mismatches; it is worth noting that the mismatches should not destroy the sense of comparison. The first version has a rigorous solution for a long time; however, new algorithms appear effective, e. g. for very long sequences or many entities under comparison [7, 8].

The second version is much more complicated since it requires a rigorous and unambiguous description of admissible mismatches and their metrization. The latter is a problem itself; currently, the alignment methodology based on the Levenshtein distance [3–5] is the most widely spread approach here. This methodology has several disadvantages, and the most crucial among them is the arbitrariness in choosing the score function, resulting in the appearance of various free parameters, making the alignment an art rather than a science; the alignment divergence since the search procedure never stops, if no special efforts are made; and finally the insertion and deletion mismatches bringing the most problematic trouble for the alignment.

V.V. Shaidurov in [6] proposed a new method of the common subsequence search which is very easy using insertions and deletions (further referred to as the Shaidurov's method). It is an alignment-free method, which is free from any adjustable parameters heavily affecting the final result of the

SibDATA 2021: The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021, June 25, 2021, Krasnoyarsk, Russia
EMAIL: annamo@icm.krasn.ru (A. Molyavko); e.d.karepova@icm.krasn.ru (E. Karepova); msad@icm.krasn.ru (M. Sadovsky)
ORCID: 0000−0003−1016−0131 (A. Molyavko); 0000−0002−6515−2932 (E. Karepova); 0000−0002−1807−0715 (M. Sadovsky); 0000−0002−0861−0001 (I. Borovikov); 0000−0001−9264−859X (O. Mutovina)

comparison. This paper shows the feasibility of the method with random four-letter symbol sequences generated by the Bernoulli process.

The method is based on calculating the convolution of two polynomials; each polynomial represents the corresponding symbol sequence. The algorithm of the method results in two sequences from the same alphabet $\aleph\{A, C, G, T\}$ of the capacity $K = |\aleph| = 4$ of the length $N1$ and $N2$, correspondingly. The algorithm comprises four steps. Let us consider the former for the comparing two sequences T1 and T2; $|T1| = N1$, $|T2| = N2$.

Preprocessing First of all, the sequence (T2, for certainty) must be inverted: $b1, b2, \ldots, bN2 \Rightarrow bN2, bN2-1, \ldots, b1$. Then, both sequences must be converted into K binary ones, $K = |\aleph|$ so that each of K binary sequences corresponds to a peculiar symbol from $\aleph$. For example, the binary sequence for G has unity instead of G and zero instead of all the other symbols. Finally, each of $4 \times 4 = 8$ binary sequences must be extended to $\widehat{L}$ which is the nearest upper power of 2 of the value $|T1| + |T2| - 1$; this expansion is required for the fast Fourier transform (FFT) implementation;

Processing Then, each of the $2K$ binary sequences is transformed through the calculation of the Fast Fourier transformation, thus changing the binary $(0, 1)$ - sequence into a complex one;

Convolution Each couple of the sequences corresponding to the same symbol must be multiplied, element by element, and the sum of the products obtained for all the symbols must be calculated, and

Inverse FFT The sum obtained at the previous step must be processed using inverse FFT which yields a real number sequence.

Each element of this real (positive) number sequence derived due to inverse FFT indicates the number of the coinciding symbols (within the overlapping pattern of these two sequences) regardless of the particular location of the coinciding symbols. This paper aims at illustrating the highly efficient feasibility of Shaidurov's method for the problem of pattern search for considering the biological matter. The patterns to be found are relatively short (70 to 600 symbols long) subsequences of high biological value, the so-called *transposons*. We search for these patterns in chloroplast genomes of various plants, including Hymnosperms and flowering plants; the point is that some researchers believe that there are no transposons in chloroplasts [9, 1].

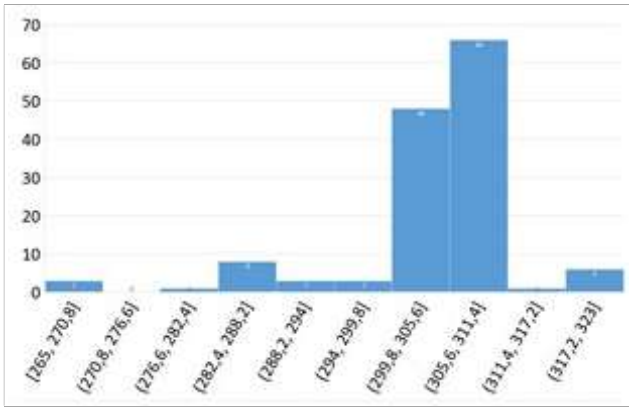
Figure 1: An example of a fuzzy coincidence found using the Shaidurov's method but with a CENSOR failure in the search

2. Results

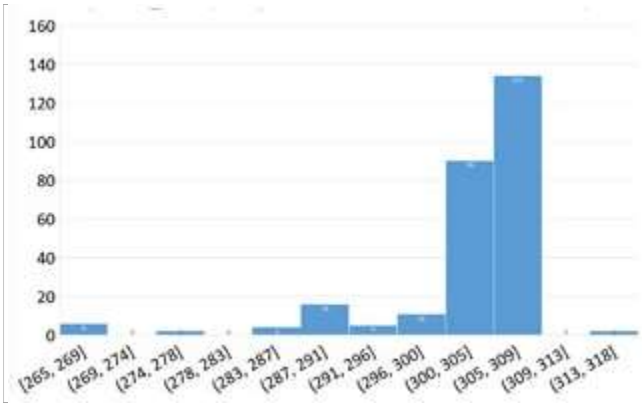
The alphabet $\aleph$ = {A, C, G, T} corresponds to genetic entities. All the pattern subsequences were retrieved from the Repbase deposit. We compared the Shaidurov's method with the standard alignment provided by the CENSOR software [2]; two parameter sets were used. Both sets are implemented in the software and support either "rigid" or "flexible" search. The Shaidurov's method is free from this arbitrariness.

CENSOR found a transposon of Copia-18 BD-I type in 181 genomes, and a transposon of MuDR-64 OS type in 44 genomes. Reciprocally, the Shaidurov's method revealed 323 and 213 entries, respectively, and these sets include those provided by CENSOR. It is evident that the Shaidurov's method is more effective than the alignment, since it identified a number of entries omitted by CENSOR; Fig. 1 illustrates this point. Here, the upper sequence corresponds to the genome, and the lower one corresponds to the transposon.

The advantages provided by the Shaidurov's method are shown in Figure 2(a) and 2(b). Figure 2(a) shows the Shaidurov's method output for the Copia-18 BD-I (length is 319 symbols) search with respect to point mutations with no insertions or deletions. Figure 2(b) shows similar results with the insertion/deletion incorporation.



a)



b)

Figure 2: An example of a fuzzy coincidence found using the Shaidurov's method but with a CENSOR failure in the search

The X axis shows the number of genomes with the pattern found in them and falling into the corresponding interval. The convolution calculation supports the identification of the transposon embedding both with and without insertions or deletions. The identification of the transposon pattern without insertions and/or deletions allows one to determine the exact number of perfectly matching symbols equal to the transposon length. The inclusion of insertions or deletions into the pattern search worsens the estimation of the exactly matching symbols, and the former becomes a probabilistic one.

The estimation goes down as the number of insertions/deletions grows up. Thus, the X axis in Fig. 2(b) shows the intervals of estimating the number of coincidences instead of the exact interval values. Fig. 2(a) unambiguously shows that Copia-18 BD-I is sure to be found in two genomes, since there are 316 coincidences among 319 ones. On the other hand, this figure shows that the greatest number of genomes (134 entities) has 305 to 309 coincidences among 319.

The results presented above unambiguously prove the presence of transposons in chloroplast genomes; however, their abundance and diversity is significantly lower in comparison with other genetic systems. Here, we are not able to make an immediate comparison of the alignment efficiency as opposed to the Shaidurov's method since the CENSOR output yields the so-called score (which is a numeric evaluation of the "quality" of the alignment), which is inspired by users from the view point of the rule for its derivation.

The results of this paper prove the high efficiency of the Shaidurov's method in the analysis of the biologically inspired comparison of symbol sequences and pattern search including insertions and deletions. Apparently, the greatest advantage of the Shaidurov's method is the complete absence of hidden or free parameters to be used to adjust the comparison.

Two problems make an essential obstacle in broader applications of the method. The former is signal/noise ratio improvement, and the latter is the localization of the sites of interest. Surprisingly, both problems could be significantly addressed with the expansion of the alphabet. Initially, we used a single letter alphabet to derive the binary sequences (that is, four letters, in the case of nucleotide sequences). However, one can assign the duplets, triplets, etc., as the letters of the new extended alphabet, doing the same job. Of course, it yields an exponential growth of the number of binary sequences to be processed. Meanwhile, they could be processed in parallel, so it is not a problem. Since the convolution value represents the number of exactly matching symbols throughout the overlap of two sequences regardless of the exact location of the matched symbols, then the expansion of the alphabet will result in a significant decrease of the number of exactly matching *k*-tipples, thus providing the better signal/noise ratio and the localization of the longer highly similar subsequences in the compared entities.

3. Conclusion

The paper presents a new method for comparison and analysis of symbol sequences. The method is based on the convolution function calculation defined over the binary numeric sequences derived from the original symbol sequence. The method provides highly parallel implementation and is very powerful in insertion or deletion mutations search. A discrete fast Fourier transform is implemented for convolution calculation. Also, an idea of the alphabet expansion is proposed to improve the signal/noise ratio. Some genomic applications are provided and discussed. The applications are used to illustrate and overcome the problem of signal/noise selection, and alignment localization.

4. Acknowledgements

This work is supported by the Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation in the framework of the establishment and development of regional Centers for Mathematics Research and Education (Agreement No. 075-02-2021-1384).

5. References

[1] J. L. Bennetzen, Transposable element contributions to plant gene and genome evolution, Plant molecular biology 42(1) (2000) 251–269.

[2] J. Jurka, P. Klonowski, V. Dagman, P. Pelton, Censor – a program for identification and elimination of repetitive elements from DNA sequences, Computers & chemistry 20(1) (1996) 119–121.

[3] V.I. Levenshtein, On perfect codes in deletion and insertion metric, Discrete Mathematics and Applications 2(3) (1992) 241–258.
[4] V.I. Levenshtein, Efficient reconstruction of sequences from their subsequences or supersequences, Journal of Combinatorial Theory, Series A 93(2) (2001) 310–332.
[5] V.I. Levenshtein, Bounds for deletion/insertion correcting codes, in: Proceedings IEEE International Symposium on Information Theory., 2002, p. 370.
[6] A. Molyavko, V. Shaidurov, E. Karepova, M. Sadovsky: Highly parallel convolution method to compare DNA sequences with enforced in/del and mutation tolerance, in: International Work-Conference on Bioinformatics and Biomedical Engineering, Springer, 2020, pp. 472–481.
[7] S. P. Tsarev, M. G. Sadovsky, New error tolerant method for search of long repeats in DNA sequences, in: International Conference on Algorithms for Computational Biology, Springer, 2016, pp. 171–182.
[8] S. P. Tsarev, M. Y. Senashova, M. G. Sadovsky, Fast algorithm for vernier search of long repeats in dna sequences with bounded error density, in: International Conference on Algorithms for Computational Biology, Springer, 2018, pp. 88–99.
[9] T. Wicker, H. Gundlach, M. Spannagl, C. Uauy, P. Borrill, R.H. Ram´ırez-Gonza´lez, R. De Oliveira, K.F. Mayer, E. Paux, F. Choulet, Impact of transposable elements on genome structure and evolution in bread wheat. Genome biology 19(1) (2018) 1–18.