Imperfect palindromes depending on GC-content

Georgii Khaziev, Alexandr Seliverstov and Oleg Zverkov

Abstract. One can calculate the similarity of a given sequence to any perfect palindrome in quadratic time. We consider imperfect palindromes in DNA depending on the guanine-cytosine content. Our results are obtained with computer assisted simulation. They are useful for identifying conserved imperfect palindromes involved in genome regulation.

Introduction

Let us consider nucleotide sequences. The nucleotides $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ form two complementary pairs: $\mathrm{c}(\mathtt{A}) = \mathtt{T}$, $\mathrm{c}(\mathtt{T}) = \mathtt{A}$, $\mathrm{c}(\mathtt{C}) = \mathtt{G}$, and $\mathrm{c}(\mathtt{G}) = \mathtt{C}$. Next, c() denotes the reverse complement, i. e., $\mathrm{c}(xy) = \mathrm{c}(y)\mathrm{c}(x)$ if x or y consists of one nucleotide. So, all perfect palindromes are of the type $x\mathrm{c}(x)$, where x denotes a sequence. In particular, a sequence of odd length cannot be a perfect palindrome. There are many works devoted to the search for perfect as well as imperfect and degenerate palindromes, refer to [1, 2]. Direct repeats are also considered, refer to [3]. We consider imperfect palindromes, i. e., sequences with gaps and mismatches in some positions.

GC-content or guanine-cytosine content is the percentage of guanine (G) or cytosine (C) in a DNA sequence. For example, the average GC-content in human genomes equals 41 per cent. In regulatory DNA, a palindrome allows a transcription factor to bind as a homodimer. Regulatory palindromes are typically imperfect, refer to [4].

For two sequences x and y, let $\mathrm{dist}(x, y)$ denote the edit distance.

There is a quadratic-time algorithm that takes two sequences x and y as input and computes the optimal partition of sequence y as a concatenation $y = wz$ that minimizes the edit distance between x and the palindrome $w\mathrm{c}(w)$. The edit distance is also computed, refer to [5].

Let us denote by $|x|$ the length of x. Let us denote by $\mathrm{imp}(x)$ the ratio of the minimum edit distance to the length of the sequence:

$$\mathrm{imp}(x) = \frac{\min\{\mathrm{dist}(x, w\mathrm{c}(w)) | x = wz\}}{|x|}.$$

The ratio shows how imperfect the palindrome is. The correctness of the definition is based on the equality $\mathrm{imp}(x) = \mathrm{imp}(\mathrm{c}(x))$.

Results

For random nucleotide sequences of length 1000 with independent positions and set GC-content, the mean and standard deviation were estimated for the values of $\mathrm{imp}(x)$, refer to Table 1. For lengths above 1000, the mean value is almost independent of the sequence length.

GC-content	Mean $\mathrm{imp}(x)$	Standard deviation
0	0.147	0.0044
5	0.172	0.0054
10	0.193	0.0056
15	0.211	0.0056
20	0.225	0.0054
25	0.237	0.0052
30	0.246	0.0049
35	0.253	0.0048
40	0.258	0.0046
45	0.261	0.0044
50	0.262	0.0044

Table 1. The empirical estimation of the mean and standard deviation of $\mathrm{imp}(x)$ for long sequences with $|x| = 1000$.

The next series of empirical estimations are performed for random sequences of length 100. It can be seen that for almost all sequences the value of $\mathrm{imp}(x)$ is close to the median, refer to Table 2.

Conclusion

We have improved known algorithms to solve some problems arising in bioinformatics. There are many examples of nucleotide sequences with perfect as well as imperfect palindromes. Our empirical estimations are useful for finding imperfect palindromes in DNA.

Funding. The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation for IITP RAS.

GC-content	0.001%	0.01%	0.1%	1%	10%	50%
0	0.08	0.09	0.10	0.12	0.14	0.16
10	0.11	0.12	0.14	0.15	0.18	0.21
20	0.14	0.15	0.17	0.19	0.21	0.24
30	0.17	0.18	0.19	0.21	0.23	0.26
40	0.18	0.19	0.21	0.22	0.25	0.27
50	0.18	0.20	0.21	0.23	0.25	0.28

TABLE 2. The empirical estimates of quantiles of $\mathrm{imp}(x)$ for short sequences with $|x| = 100$.

References

[1] Alzamel M., Hampson C., Iliopoulos C.S., Lim Z., Pissis S., Vlachakis D., Watts S. Maximal degenerate palindromes with gaps and mismatches. *Theoretical Computer Science.* 2023, vol. 978, article no. 114182. https://doi.org/10.1016/j.tcs.2023.114182

[2] Mieno T., Funakoshi M., Nakashima Y., Inenaga S., Bannai H., Takeda M. Computing maximal palindromes in non-standard matching models. *Information and Computation.* 2025, vol. 304, article no. 105283. https://doi.org/10.1016/j.ic.2025.105283

[3] Lafond M., Lai W., Liyanage A., Zhu B. The longest subsequence-duplicated subsequence and related problems. *Information and Computation.* 2025, vol. 306, article no. 105313. https://doi.org/10.1016/j.ic.2025.105313

[4] Datta R.R., Rister J. The power of the (imperfect) palindrome: sequence-specific roles of palindromic motifs in gene regulation. *Bioessays.* 2022, vol. 44, no. 4, article no. e2100191. https://doi.org/10.1002/bies.202100191

[5] Zverkov O., Seliverstov A., Shilovsky G. Alignment of a hidden palindrome. *Mathematical Biology and Bioinformatics.* 2024, vol. 19, no. 2, pp. 427–438. (In Russian.) https://doi.org/10.17537/2024.19.427

Georgii Khaziev
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
e-mail: khaziev@iitp.ru

Alexandr Seliverstov
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
e-mail: slvstv@iitp.ru

Oleg Zverkov
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
e-mail: zverkov@iitp.ru