Imperfect palindromes depending on GC-content

Georgii A. Khaziev[1] Alexandr V. Seliverstov[1] Oleg A. Zverkov[1]

[1]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow

July 15, 2025

Introduction. Palindromes in Bioinformatics.

Let us consider nucleotide sequences. The nucleotides $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ form two complementary pairs:

$$\mathrm{c}(\mathtt{A}) = \mathtt{T},\ \mathrm{c}(\mathtt{T}) = \mathtt{A},\ \mathrm{c}(\mathtt{C}) = \mathtt{G},\ \text{and}\ \mathrm{c}(\mathtt{G}) = \mathtt{C}.$$

Next, $\mathrm{c}()$ denotes the reverse complement, i. e., $\mathrm{c}(xy) = \mathrm{c}(y)\mathrm{c}(x)$. For example, $\mathrm{c}(\mathtt{AACG}) = \mathtt{CGTT}$. In DNA, a perfect palindrome is an inverted sequence repeat, i. e., reverse complement of itself. Let us omit the concatenation symbol. 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 any perfect palindrome. There are many examples of nucleotide sequences with perfect as well as imperfect palindromes [Zverkov et al., 2024].

Introduction.GC content

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 [Datta and Rister, 2022].

Results. imp function

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

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

Example

For $x = \mathrm{ATATGT}$, $\mathrm{imp}(x) = \frac{1}{6}$

The ratio shows how imperfect the palindrome is. The correctness of the definition is based on the next theorem.

Results. Theorem 1

Theorem

There is a quadratic-time algorithm that computes the $\mathrm{imp}(x)$ *function as well as optimal partition of the input sequence* x *as a concatenation* $x = wz$ *that minimizes the edit distance between* x *and the palindrome* $w\mathrm{c}(w)$.

Results. Theorem 2

Theorem

$\mathrm{imp}(x) = \mathrm{imp}(\mathrm{c}(x))$.

Results. Theorem 3

Theorem

For all even-length sequences it is true that $\mathrm{imp}(x) \leq 1/2$. *For all odd-length sequences it is true that* $0 < \mathrm{imp}(x) \leq (1 + 1/|x|)/2$.

Results.

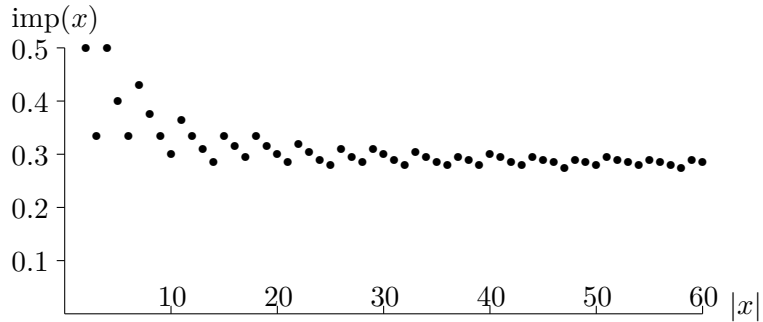

Figure: Empirical dependency between median of $\mathrm{imp}(x)$ and $|x|$ for random nucleotide sequences for $gc = 0.5$.

Results.

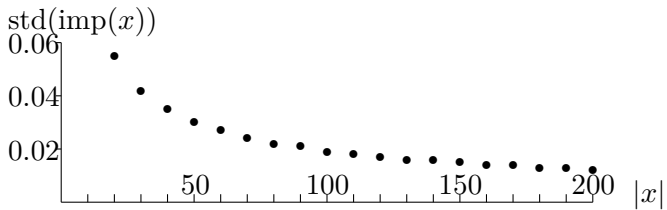

Figure: Empirical dependency between standard deviation of $\mathrm{imp}(x)$ and $|x|$ for random nucleotide sequences for $gc = 0.5$.

Results.

For $gc = 0.5$ standart deviation of $\mathrm{imp}(x)$ have approximation with determination coefficient $R^2 = 0.999$:

$$std_{0.5} = \frac{(|x| - 0.6)^{0.35}}{2.58|x|}.$$

For $gc = 0$ standart deviation of $\mathrm{imp}(x)$ have approximation with determination coefficient $R^2 = 0.999$:

$$std_0 = \frac{(|x| - 0.1)^{0.36}}{2.65|x|}.$$

Results.

gc	0.001%	0.01%	0.1%	1%	10%	50%	std
0.00	0.08	0.09	0.10	0.12	0.14	0.16	0.020
0.05	0.09	0.11	0.12	0.14	0.16	0.18	0.022
0.10	0.11	0.12	0.14	0.15	0.18	0.21	0.022
0.15	0.13	0.14	0.15	0.17	0.19	0.22	0.022
0.20	0.14	0.15	0.17	0.19	0.21	0.24	0.022
0.25	0.15	0.17	0.18	0.20	0.22	0.25	0.021
0.30	0.17	0.18	0.19	0.21	0.23	0.26	0.021
0.35	0.17	0.18	0.20	0.22	0.24	0.27	0.020
0.40	0.18	0.19	0.21	0.22	0.25	0.27	0.020
0.45	0.19	0.19	0.21	0.23	0.25	0.27	0.019
0.50	0.18	0.20	0.21	0.23	0.25	0.28	0.019

Table: Empirical quantiles for $\mathrm{imp}(x)$ and standart deviation for $|x| = 100$.

Results.

gc	$mean$	std
0.00	0.147	0.0044
0.05	0.172	0.0054
0.10	0.193	0.0056
0.15	0.211	0.0056
0.20	0.225	0.0054
0.25	0.237	0.0052
0.30	0.246	0.0049
0.35	0.253	0.0048
0.40	0.258	0.0046
0.45	0.261	0.0044
0.50	0.262	0.0044

Table: Empirical estimation for mean and standart deviation for $\mathrm{imp}(x)$ and sequences with length greater or equal to 1000.

Results.

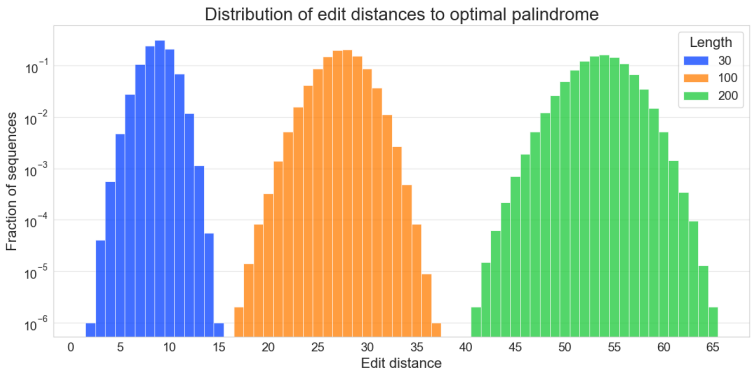

Figure: Distribution of edit distances to optimal palindrome for 1000000 random sequences with $gc = 0.5$

Results

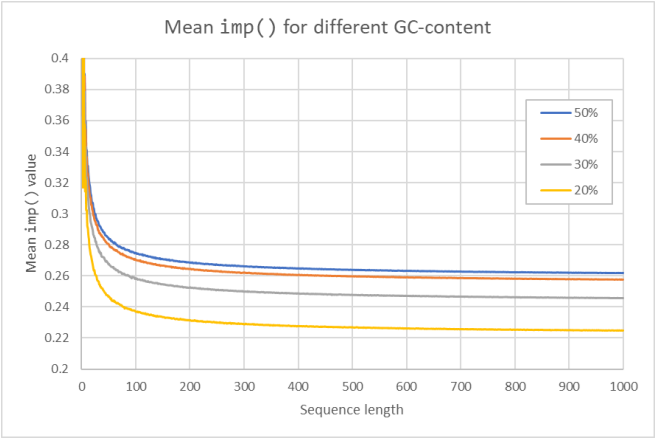

Figure: Dependencies between mean of $\mathrm{imp}(x)$ and $|x|$ for different GC content.

Results. trimmers

Substring is a contiguous sequence of characters within a string. The main idea behind the algorithm to select an imperfect palindrome is checking whether one of the optimal lengths of the prefix w of the input sequence x differs significantly from $|x|/2$. If such case occurs, then the algorithm deletes either prefix or suffix by difference between $|w|$ and $|x|/2$.

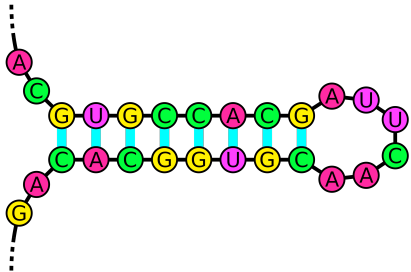

Figure: Hairpin example.

Results. trimmers

All three functions `pref_trimmer`, `suff_trimmer`, and `double_trimmer` take as input nucleotide sequence x, sorted list `optimal_lengths` of optimal prefix lengths $|w|$, and floating-point number `cutoff_condition`. Note that $\min(\texttt{optimal\_lengths})$ and $\max(\texttt{optimal\_lengths})$ are the first and last elements of `optimal_lengths`, respectively.

Results.

The `pref_trimmer` function trims first rd symbols of x, where

$$rd = \max(\texttt{optimal\_lengths}) - \left\lfloor \frac{|x|}{2} \right\rfloor,$$

when $rd \geq \text{length}(x) \cdot$ `cutoff_condition` is satisfied.
The `suff_trimmer` function trims last ld symbols of x, where

$$ld = \left\lfloor \frac{|x|}{2} \right\rfloor - \min(\texttt{optimal\_lengths}),$$

when $ld \geq |x| \cdot$ `cutoff_condition` is satisfied.

Results. double_trimmer

The `double_trimmer` function initially computes

$$rd = \max(\texttt{optimal\_lengths}) - \left\lfloor \frac{|x|}{2} \right\rfloor$$

and

$$ld = \left\lfloor \frac{|x|}{2} \right\rfloor - \min(\texttt{optimal\_lengths}).$$

Subsequently it checks if

$$\begin{cases} rd \geq |x| \cdot \texttt{cutoff\_condition} \\ ld \geq |x| \cdot \texttt{cutoff\_condition} \end{cases}$$

If the first inequality is satisfied, the function trims the first rd symbols from the string x. Similarly, if the second inequality is satisfied, the function trims the last ld symbols from x.

Theorem

For perfect palindrome x for any `cutoff_condition` > 0 *no trimming would be performed.*

Theorem

For any $n \geq 3$, there exist `cutoff_condition` > 0 *and x with length of at least n, which satisfies both prefix and suffix trimming conditions such that*

$$\texttt{pref\_trimmer}(x_{suff}) \neq \texttt{suff\_trimmer}(x_{pref}),$$

where x_{suff} and x_{pref} are results of suffix and prefix trimming of x, respectively.

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. In particular, it is crucial for predicting gene expression regulations as well as RNA structures. The implementation of algorithms in Python will enable a wide range of bioinformaticians to apply them in their work. Moreover, low computational complexity allows efficient processing of large datasets.

References


Datta, R. R. and Rister, J. (2022).
The power of the (imperfect) palindrome: Sequence-specific roles of palindromic motifs in gene regulation.
BioEssays, 44(4):2100191.


Zverkov, O. A., Seliverstov, A. V., and Shilovsky, G. A. (2024).
Alignment of a hidden palindrome.
Mathematical Biology and Bioinformatics, 19(2):427–438.