Efficient Storage of Similar Sequences Using Elastic-Degenerate Strings

Dominika Bohuslavová

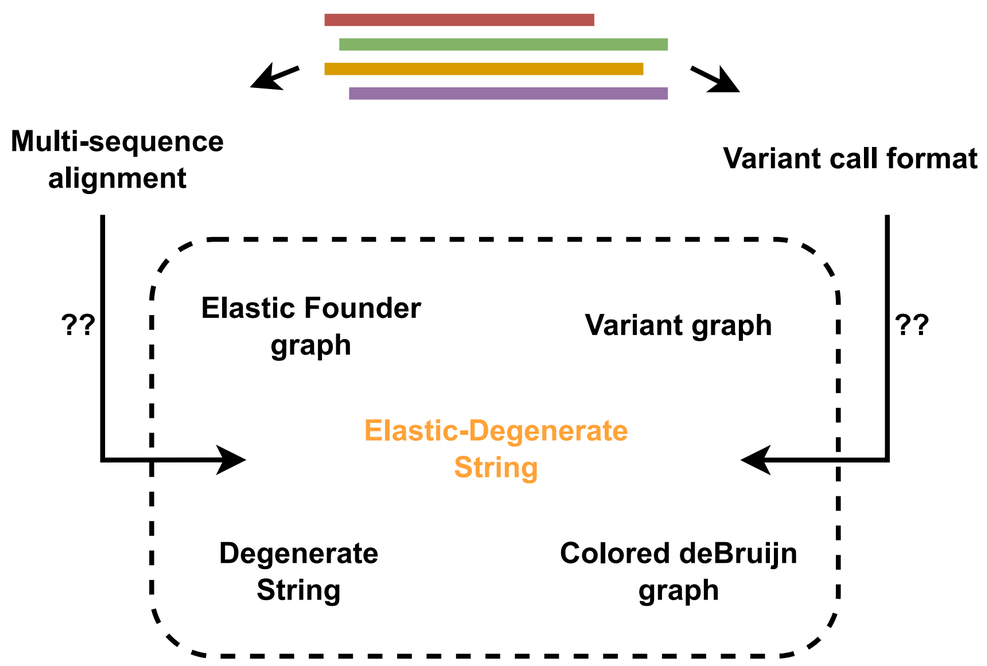
Faculty of Information Technology, Czech Technical University in Prague, Prague, Czech Republic

Introduction

As the volume of genomic data grows, effective data management and efficient storage are key challenges in bioinformatics. However, these data also reveal similarities and redundancies between sequences, which can lead to good compression potential. A promising representation of such data that exploits its compression potential is the **Elastic-Degenerate String (EDS)**, which efficiently stores set of sequences in a compact format.

EDS not only allows effective compression of genomic information but also has the potential to predict future data of the same type. EDS allows incoming sequences to be integrated into it's information capacity, thereby improving our understanding of genomic variability.

Despite these advantages, several open problems remain. A major challenge is the creation of EDS from a given dataset, as we aim to maximize the accurate reflection of the dataset's meaning while minimizing the size of the EDS. Additionally, the development of efficient data structures for querying such strings is essential for their practical application in bioinformatics.



Elastic-degenerate string

$$X = \{AAC\}\left\{\begin{matrix}\varepsilon\\TA\end{matrix}\right\}\left\{\begin{matrix}A\\G\\T\end{matrix}\right\}\{GAC\}\left\{\begin{matrix}A\\GG\end{matrix}\right\}\{AA\},$$

$n = 6,$
$N = 3+3+3+3+3+2 = 17,$
$m = 1+2+3+1+2+1 = 10,$
$L(X) = \{AACAGACTGAAA, AACGGACTGAAA, \ldots\}$

I-EDS, context length $l = 5$

blocks: C_0 R_0 C_1 R_1 C_2

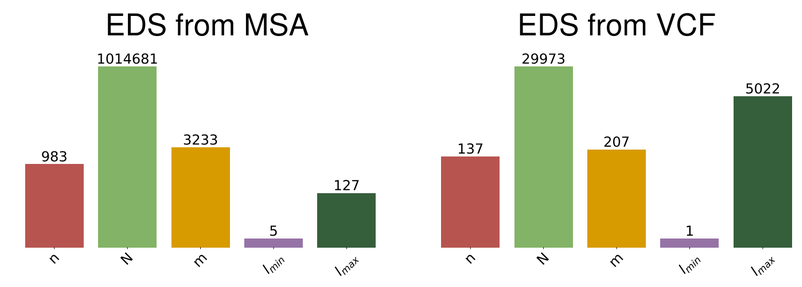
$$Y = \{ACGTAC\}\left\{\begin{matrix}\varepsilon\\ACG\end{matrix}\right\}\{CAGCA\}\left\{\begin{matrix}T\\GGAG\end{matrix}\right\}\{ATT\}$$

Definition and metrics

- An **elastic-degenerate string** X is a string, where $S[i]$, for $0 \leq i \leq n$ is a **nonempty** set of different strings.
- The size $N = ||X||$ is the total number of characters in the degenerate string including ε.
- An empty string ε is counted as one symbol for each j such that $\varepsilon \in S[j]$.
- We also denote symbols $S[i]$, where $|S[i]| > 1$ is **degenerate symbol**.
- For elastic-degenerate strings we define $m = \sum_{i=1}^{n} |S[i]|$ and the **language of EDS** $L(X)$, is the set of all strings $s = s_1 s_2 \ldots s_n$ such that $s_i \in S[i]$ for all $1 \leq i \leq n$.
- As a **context length** l we mean the length of non-degenerate symbol.
- A **phased EDS** is such X', where its language $L(X') \subset L(X)$. It stores additional information about allowed paths in EDS.

SARS-CoV-2 [3]	
Reference FASTA 33kB	Variants FASTA 9MB

VCF (ref + .vcf)	$\approx$ 50kB $\Longrightarrow$	EDS	$\approx$ 33kB*
MSA	$\approx$ 11MB $\Longrightarrow$	EDS	$\approx$ 2MB



* Depends on the number of variants, here it is minor compared to the reference size.

Open problems

- How and for what kind of data is it suitable to create EDS? Pangenomes, smaller sets of highly similar data, or perhaps something else?
- What is the ideal balance between preserving the information from the original collection, reducing the space, and keeping predictive capabilities of EDS? [1, 2]
- What can we do with such a structure? Indexing, exact and approximate pattern searching, further compression?
- What are the advantages compared to other structures that can similarly encode a collection of sequences (Variant graphs, . . .)?

References

[1] Aleksander Cisłak and Szymon Grabowski. *SOPanG 2: online searching over a pan-genome without false positives*. Apr. 2020. DOI: 10.48550/arXiv.2004.03033.

[2] Esteban Gabory et al. "Comparing elastic-degenerate strings: Algorithms, lower bounds, and applications". In: *34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023)*. 2023.

[3] James Hadfield et al. "Nextstrain: real-time tracking of pathogen evolution". In: *Bioinformatics* 34.23 (May 2018), pp. 4121–4123. DOI: 10.1093/bioinformatics/bty407.

[4] Pavel Sagulenko, Vadim Puller, and Richard A Neher. "TreeTime: Maximum-likelihood phylodynamic analysis". en. In: *Virus Evol* 4.1 (Jan. 2018), vex042.

Acknowledgements

This work is supported by the Grant Agency of the Czech Technical University in Prague grant No. SGS23/205/OHK3/3T/18 and by ROBOPROX project reg. no. CZ.02.01.01/00/22_008/0004590.

For more information . . . https://draessld.github.io/mysite/