Computational Pan-Genomics with Elastic-Degenerate Strings

(a case study of my research)

NADIA PISANTI
(this Department)

The pan-Genome

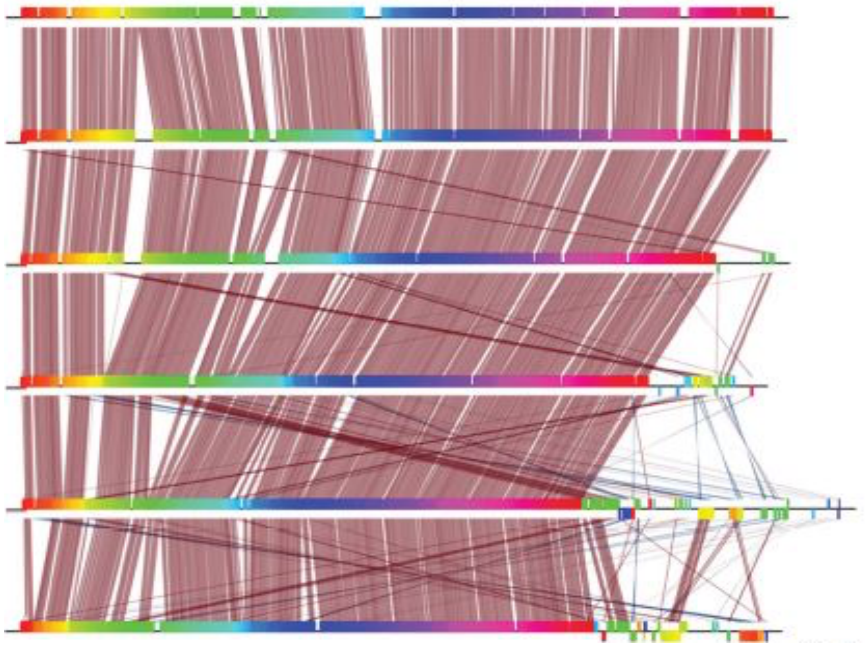
Some definitions of pan-genome:

- *... describes the full complement of genes [...] which can have large variation in gene content among closely related strains* [Wikipedia]
- *a collection of genomic sequences to be analyzed jointly or to be used as a reference*

[The Computational Pan-Genomics Consortium, 2016]

Traditionally, a reference genome is:

- a genome of a single selected individual, or
- a consensus drawn from a population, or
- a "functional" genome, or
- a maximal genome capturing all ever-detected sequences
- ...



ED-strings

Sequence Alignment:

Reference genome: ...ATGCAACGGGTA--TTTTA...

Individual 1 read: ATGCAACGGGTATATTTTA

Individual 2 read: ATGCACCTGG----TTTTA

Representation (Huang et al, Bioinformatics, 2013):

$$\{\ \texttt{ATGCA}\ \} \cdot \left\{ \begin{array}{c} \texttt{A} \\ \texttt{C} \end{array} \right\} \cdot \{\ \texttt{C}\ \} \cdot \left\{ \begin{array}{c} \texttt{G} \\ \texttt{T} \end{array} \right\} \cdot \{\ \texttt{GG}\ \} \cdot \left\{ \begin{array}{c} \texttt{TA} \\ \texttt{TATA} \\ \varepsilon \end{array} \right\} \cdot \{\ \texttt{TTTTA}\ \}$$

Elastic Degenerate string as a natural representation of a pan-genome

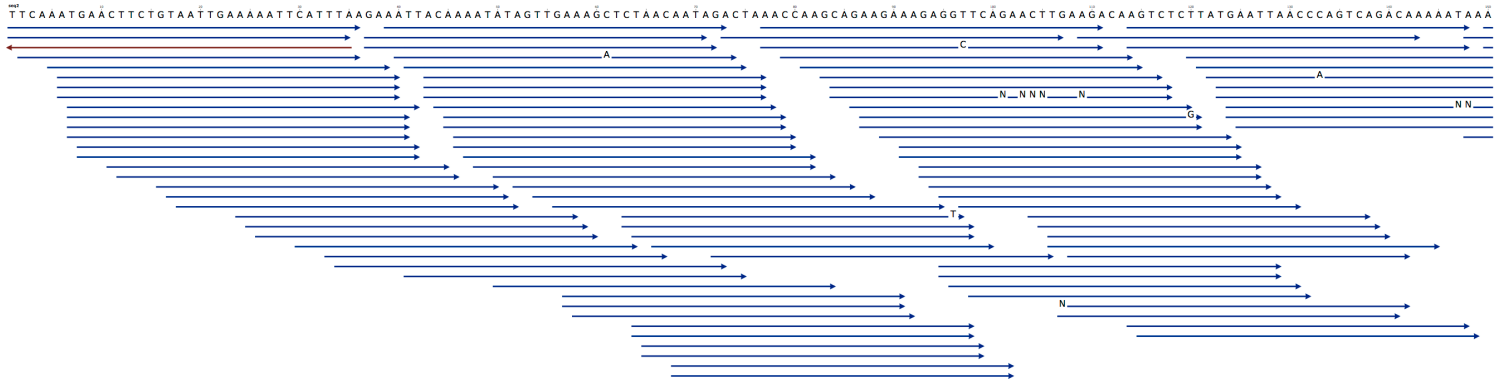
It corresponds to the Variant Call Format (.vcf) standard
[e.g. data from the 1000 Genomes project]

Reference Pan-Genome

- Cheaper sequencing: re-sequencing became a common task.

- In genome analysis workflows, downstream of re-sequences there is the task of

mapping reads (a string) on a reference genome (a longer string)
It's PATTERN MATCHING: read is P, reference genome is T



EDSM problem

ELASTIC DEGENERATE STRING MATCHING (EDSM)

Input: a string P of length m, an ED string $\tilde{T}$ of length n and total size N

Output: all positions in $\tilde{T}$ where at least one occurrence of P ends

P = CGGGTATA

$$\tilde{X} = \{ \text{TTC} \} \cdot \left\{ \begin{matrix} \text{G} \\ \text{T} \end{matrix} \right\} \cdot \{ \text{GG} \} \cdot \left\{ \begin{matrix} \text{TA} \\ \text{TATAG} \\ \varepsilon \end{matrix} \right\}$$

Lower bounds & upper bounds

[ICALP 2019]

In [CPM 2017] we solved EDSM in $O(N + n*\mathbf{m^2})$ time

In [CPM 2018] they solve it in $O(N + n*\mathbf{m^{1.5}\sqrt{(\log m)}})$ time

Can EDSM be improved further?

In [ICALP 2019] we solve EDSM in $O(N + n*\mathbf{m^{1.381}})$ time

... with an algebraic method!

We show one can't do better with combinatorial methods

Pattern Matching on ED-string with errors

[SPIRE 2017]

Reads carry sequencing errors: how can we represent them?

Hamming Distance:

Given two strings X and Y on the same alphabet and having the same length, the *Hamming Distance* $d_H(X,Y)$ between X and Y is the number of positions in which they differ.

X = CGGGTATA
Y = CAGGCATA

$d_H(X,Y)=2$

Edit Distance:

Given two strings X and Y on the same alphabet, the *edit Distance* $d_E(X,Y)$ is the number of substitutions, insertions, or deletion of a letter needed to transform X into Y (or viceversa, as $d_E(X,Y)=d_E(Y,X)$).

X = CGGGTAT--A
Y = CCGG--ATTA

$d_E(X,Y)=3$

Degenerate Strings Comparison

STRING COMPARISON among (E)D-strings

is a basic tool for many other problems:

Are two degenerate strings the same? Or similar?

Or share sub-(E)D-strings? Motifs?

Is one (E)D-string a substring of another (E)D-string?

A Reverse? A Palindrome?

Degenerate Strings Comparison

our result [WABI 2018]

A definition of a match among D-strings

(a step into formal languages and automata problems)

A linear (O(N+M)) algorithm to tell whether two D-strings X (of size N) and Y (of size N) do match

("accidentally" solving an open formal languages and automata problem)

An application of such D-strings comparison to the design of two algorithms to decompose a D-string into palindromes

(a proof-of-concept on real RNA data)

References

The Computational Pan-Genomics Consortium:
Computational pan-genomics: status, promises and challenges.
Briefings in Bioinformatics 19(1): 118-135 (2018)

R.Grossi, C.S.Iliopoulos, C.Liu, N.Pisanti, S.P. Pissis, A.Retha, G.Rosone, F.Vayani, L.Versari:
On-Line Pattern Matching on Similar Texts.
CPM 2017: 9:1-9:14

G.Bernardini, N.Pisanti, S.P. Pissis, G.Rosone:
Pattern Matching on Elastic-Degenerate Text with Errors.
SPIRE 2017: 74-90
[extended version in press in Theoretical Computer Science journal]

M.Alzamel, L.A.K. Ayad, G.Bernardini, R.Grossi, C.S.Iliopoulos, N.Pisanti, S.P.Pissis, G.Rosone:
Degenerate String Comparison and Applications.
WABI 2018: 21:1-21:14

G.Bernardini, P.Gawrychowski, N.Pisanti, S.P.Pissis, G.Rosone:
Even Faster Elastic-Degenerate String Matching via Fast Matrix Multiplication.
ICALP 2019: 21:1-21:15