

Simplified pangenome graph traversals with PSSM scoring : search for motifs differentials

Julien Guidihounme, Simon De Givry, Benjamin Linard

▶ **To cite this version:**

Julien Guidihounme, Simon De Givry, Benjamin Linard. Simplified pangenome graph traversals with PSSM scoring : search for motifs differentials. JOBIM 2024, Jun 2024, Toulouse, France. ⟨hal-04656256⟩

HAL Id: hal-04656256

https://hal.science/hal-04656256v1

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Simplified pangenome graph traversals with PSSM scoring : search for motifs differentials

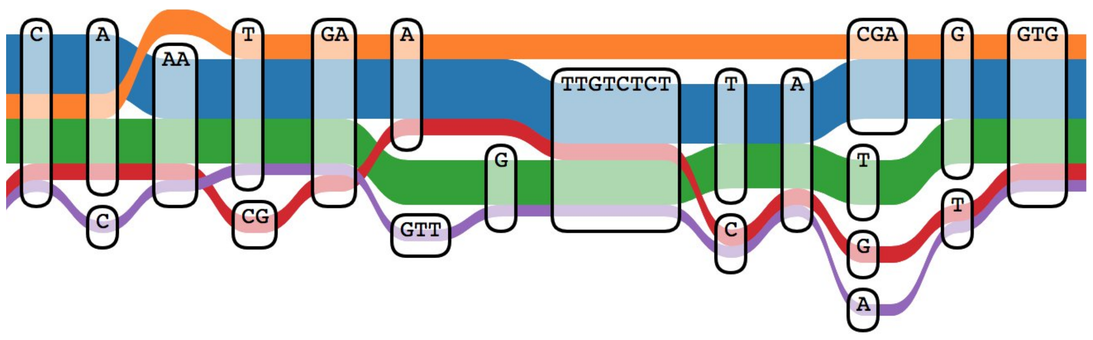
Julien GUIDIHOUNME[1], Simon DE GIVRY[1] and Benjamin LINARD[1]

1 Unité de Mathématiques et Informatique Appliquées, INRAE,Castanet-Tolosan, France
Corresponding Author: julien.guidihounme@inrae.fr

CONTEXT

A pangenome Variation Graph (VG) is a data structure that integrates complete information about genetic variations in a species or group of species [1]. Computational pan-genomic is beginning to be discussed as a viable alternative to many standard approaches previously based on sequences [2]. Although in itself a compression of genome sequences, a fundamental problem is the size of this graph (tens to hundreds of millions of nodes): the application of numerous algorithms based on traversals is not possible and heuristics and indexing techniques are necessary to be able to manipulate and query it.
We are interested in developing an approach to search for functional differentials in a VG via the detection of sequence pattern changes between the genomes included in the graph. To do so, we search for probabilistic pattern described via a PSSM (Position Specific Scoring Matrix)[3]. Our approach aims to output all significant PSSM score differentials that can be extracted from a VG. Note we detect motifs on node boundaries only (the score difference is either null or undefined for motifs inside nodes).

Overview of a pangenome graph [4]



Functionnal motifs described via a PSSM :

Patterns from alignment

GAGGT
TCCGT
CAGGT
ACAGT
TAGGT

PSSM

A	-0.32	1.26	-0.32	$-\infty$	$-\infty$
C	-0.32	0.67	-0.32	$-\infty$	$-\infty$
G	-0.32	$-\infty$	1.26	2	$-\infty$
T	0.67	$-\infty$	$-\infty$	$-\infty$	2



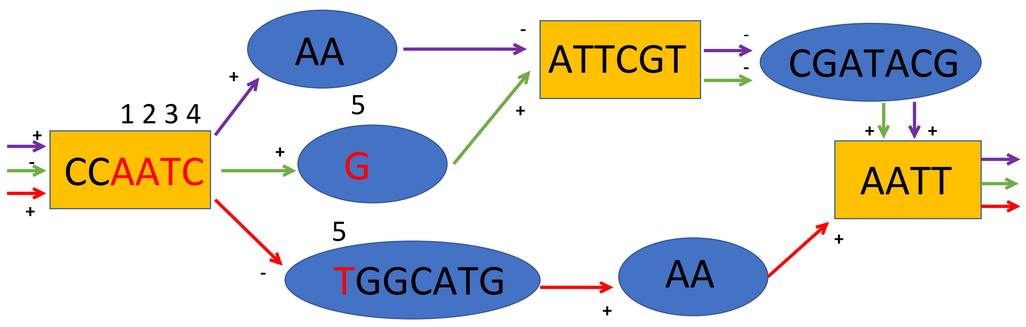
a. Determine the score associated with the motif (PSSM) in the graph

b. Find for the differential within score the genomic species integrated into the graph

c. Scale problem: whence our method on subgraphs defined by pair of genomes to find the differential score

METHOD OVERVIEW

- Our approach iterates on genome **pairs**
- Below example shows 3 genomes (3 possible pairs)



- **Yellow nodes** : for each pair, **the "bifurcation" nodes** are listed, they are the starting point to define sequence positions allowing a putative differential
- **In red text:** for the pair (green, red) and for a PSSM matrix of length M, positions where scoring is necessary are selected.
- **Scoring:** We use classic PSSM scoring algorithms[5][6] to detect a differential. Positions in the bifurcation node are computed once.
- **Iteration:** this process is repeated for all bifurcations, and all genome pairs

FORMALISM & IMPLEMENTATION

Definition : Intersection of genomes $I : \mathcal{P}(\mathscr{P}) \to \mathcal{P}(\mathscr{V}),\ J \mapsto \bigcup_{v \in \bigcap_{p \in J} p} \{v\}$

Property :

$$\text{Let } x \in I(\{p_i, p_j\}), i \neq j, (i,j) \in \{1,2,\dots,|\mathscr{P}|\}^2$$
$$p_i = (v_0, v_1, \dots v_{n-1}, x, v_{n+1} \dots, v_{|p_i|}) \text{ and } p_j = (s_0, s_1, \dots, s_{m-1}, x, s_{m+1}, \dots, s_{|p_j|})$$
$$\text{such that } f_i = (v_0^{i_0}, v_1^{i_1}, \dots v_{n-1}^{i_{n-1}}, x^{i_n}, v_{n+1}^{i_{n+1}} \dots, v_{|f_i|}^{i_{|f_i|}}) \text{ and } f_j = (s_0^{j_0}, s_1^{j_1}, \dots, s_{m-1}^{j_{m-1}}, x^{j_m}, s_{m+1}^{j_{m+1}}, \dots, s_{|f_j|}^{j_{|f_j|}})$$
$$\text{with } i_l \in \{+,-\}, l \in \{0,1,\dots|f_i|\} \text{ and } j_t \in \{+,-\}, t \in \{0,1,\dots|f_j|\}.$$

Then x is a bifurcation **node** if it satisfies one of the two conditions :

i . v_{n-1} *et* s_{m-1} exist, then $v_{n-1} \neq s_{m-1}$ ou $i_{n-1} \neq j_{m-1}$

ii . v_{n+1} *et* s_{m+1} exist, then $v_{n+1} \neq s_{m+1}$ ou $i_{n+1} \neq j_{m+1}$

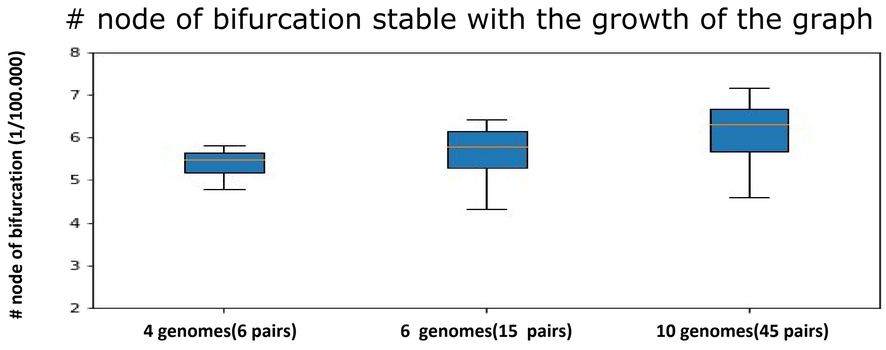
Implementation: C++ code, command-line tool

Input:
pangenome graph in GFA format + some PSSM matrix in jaspar format
Output:
a list which contains differentials: the score differential value, the index of the bifurcation node in the graph, the corresponding genome positions

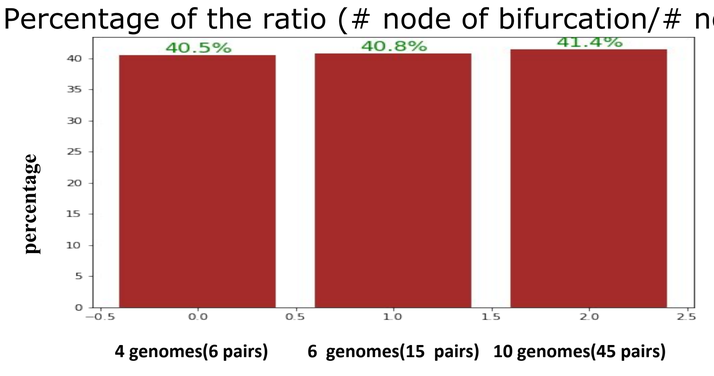
PRELIMINARY RESULTS

Tendency of bifurcation node on the VG of cabbage with 4, 6 and 10 genomes for the same chromosome. (Data courtesy of Fabrice Legai & Alexis Merguez, INRAE, Génotoul Bioinfo)

node of bifurcation stable with the growth of the graph

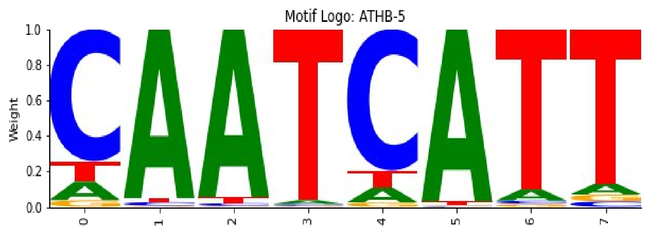


Percentage of the ratio (# node of bifurcation/# node total)



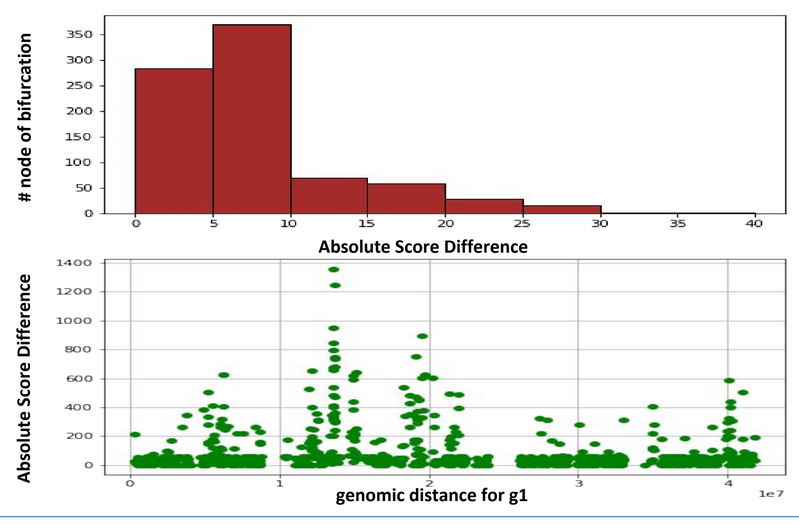
Score differentials for the ATHB-5 motif in an apricot graph : 6 genomes (genome size ~230mb), chromosome 1 only, 2210860 nodes and 3026030 edges. (Data Courtesy of Véronique Decroocq's team, INRAE, BFP, Bordeaux)

ATHB5: a regulatory motif from the jaspar database, with a role in the growth and germination of seeds (BL Raminger et al, 2023).



- Score differential for two genomes g1 and g2 in the apricot graph by setting the p-value of g1 to 10^{-4}.

- 350 regions (bifurcation nodes) show a low score differential (interval [0-10]).
- Less than 50 regions show a differential >10 , e.g. a significant motif change with putative change/loss of function.



DISCUSSION

- For a VG of one apricot chromosome, we detected 1029524 regions where we calculate score differentials (all 15 pairs of genome) in then less than 15 minutes.
- We continue to optimize the algorithm to avoid redundant computations : use of a score threshold, reuse of score computation between pairs ...
- Code needs to be extended to allow as inputs a collection of PSSM matrices.

[1] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9. [2] Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020 Aug 31;21:139-162. doi: 10.1146/annurev-genom-120219-080406. Epub 2020 May 26. PMID: 32453966; PMCID: PMC8006571. [3] Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. Nucleic Acids Res. 2009 Feb;37(3):939-44. doi: 10.1093/nar/gkn1019. Epub 2008 Dec 23. PMID: 19106141; PMCID: PMC2647310. [4] Wolfgang Beyer, Adam M Novak, Glenn Hickey, Jeffrey Chan, Vanessa Tan, Benedict Paten, Daniel R Zerbino, Sequence tube maps: making graph genomes intuitive to commuters, Bioinformatics, Volume 35, Issue 24, December 2019, Pages 5318–5320, [5] Tognon M, Bonnici V, Garrison E, Giugno R, Pinello L. GRAFIMO: Variant and haplotype aware motif scanning on pangenome graphs. PLoS Comput Biol. 2021 Sep 27;17(9):e1009444. doi: 10.1371/journal.pcbi.1009444. PMID: 34570769; PMCID: PMC8519448. [6] Touzet, H., Varré, JS. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol* **2**, 15 (2007). https://doi.org/10.1186/1748-7188-2-15