



HAL
open science

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. *Methods for Interfacing with Graphs of Genomic Sequences*, camille marchet; guillaume gautreau, Sep 2024, Lille, France. <hal-04803005>

HAL Id: hal-04803005

<https://hal.science/hal-04803005v1>

Submitted on 25 Nov 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 GUIDIHOUNME¹, Simon DE GIVRY¹, Benjamin LINARD¹

¹ Unité de Mathématiques et Informatique Appliquées, INRAE, Castanet-Tolosan, France

Corresponding author : benjamin.linard@inrae.fr

Category : Algorithm & software for pangenomics

Keywords : pangenome, variation graph, differentials, sequence motifs, PSSM

Summary :

A pangenome variation graph (VG) is a model that compresses complete genomes [1]. Nodes are labeled by genome sub-sequences and edges represent their contiguity in, at least, one genome. Original genomes sequences are modeled as colored traversals of the graph. Today, VGs are primarily utilized for analyzing structural variants in eukaryotic species. Although many tools exist to index and query a VG, these developments mainly focused on the process of sequence-to-graph mapping and subsequent variant calling. For many well-established genomic analyzes such as functional annotation or specific motifs detection, no tools facilitates straightforward querying of a VG.

We investigated the possibility to identify regulatory motif differentials from a VG when motifs are defined as a probabilistic matrices, such as a PWM or a PSSM. Existing tools allowing such analysis in a VG extract the linear sequence from the graph and apply traditional scoring schemes [2], thus duplicating many score computations and overlooking the advantages of the graph representation. We aimed to exploit the compression inherent to the graph with a targeted approach. Initially, for each pair of genomes we identify a subset of nodes with specific properties, the “bifurcation nodes”. From these nodes, we defined the chains and sequence positions for which computation of differentials is relevant. Subsequently, non-redundant computations of PSSM scores were performed, and only the most significant differences were reported (via a user-defined probability threshold, see [3]). The final output is a set of graph nodes and associated genome positions where putative functional changes may have occurred. This approach, along with the consecutive optimization problems it reveals, will be presented in details.

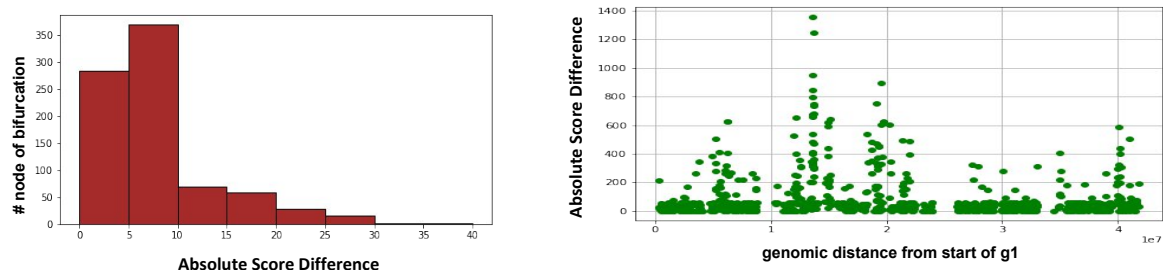


Figure: Score differentials for a genome pair (g_1 , g_2) in chromosome 1 from an apricot VG (data courtesy of V. Decrooq's team, INRAE, BFP, Bordeaux), by setting the p -value threshold of g_1 to 10^{-4} . Most regions (bifurcation nodes) are associated to a low score differential (interval [0-10]), and 50 regions show a high differential (>10), e.g. a significant motif change resulting to putative change/loss of biological function.

[1] Eizenga JM et al. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020

[2] Tognon M et al. GRAFIMO: Variant and haplotype aware motif scanning on pangenome graphs. PLoS Comput Biol. 2021

[3] Touzet, H. et al. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol*, 2007