POLITECNICO DI TORINO
Repository ISTITUZIONALE

Information theoretic clustering of the human pangenome minigraph

Original
Information theoretic clustering of the human pangenome minigraph / Ferrero, R., Gandino, F., Carbone, A.. - In: PATTERN RECOGNITION LETTERS. - ISSN 0167-8655. - STAMPA. - 191:(2025), pp. 117-123. [10.1016/j.patrec.2025.03.004]

Availability:
This version is available at: 11583/2999081 since: 2025-04-11T12:48:29Z

Publisher:
Elsevier

Published
DOI:10.1016/j.patrec.2025.03.004

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

26 November 2026

Pattern Recognition Letters 191 (2025) 117–123

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Information theoretic clustering of the human pangenome minigraph

Renato Ferrero [a], Filippo Gandino [a,*], Anna Carbone [b]

[a] Department of Control and Computer Engineering, Politecnico di Torino, Italy
[b] Department of Applied Science and Technology, Politecnico di Torino, Italy

ARTICLE INFO

Editor: Xiaoning Qian

Keywords:
Entropy measures
Natural vs Artificial Patterns Partition
Compositional self-similarity
Nucleotide patterns

ABSTRACT

Information theoretic clustering, long-range correlation, power-law scaling and self-similarity concepts have been broadly adopted for characterizing genomic features such as nucleotide composition, flexibility and bending. In this work, the 24 chromosomes of the human pangenome minigraphs, recently assembled by the Human Pangenome Reference Consortium (HPRC), are investigated to check to what extent self-similarity and scaling features are preserved in comparison to the reference linear sequences of the T2T-CHM13 individual. By taking the nucleotide self-similarity of the reference chromosomes as benchmark, it is shown that the pangenome minigraph segments exhibit lower self-similarity of the nucleotide composition compared to the linear sequence. The proposed information measures can be adopted to quantify the nucleotide self-similarity patterns and complement standard alignment techniques towards the coherent definition of the genomic profile of each species.

1. Introduction

While linear genomic sequences reflect the evolutionary complexity of single individuals, structured graphs collecting genomic fragments of multiple individuals compiled into a single data structure (*pangenome*) could feature the whole genetic profile of a given species. Graph-based approaches of branching and merging paths have been proposed to gather together features of several individuals and shed light on the evolutionary diversity of each species [1]. The definition of the optimal pangenome architecture is a challenging endeavour of increasing interest [2–4], which could benefit from the interdisciplinary perspective provided by the computational and complex system's approaches to pattern analysis.

Several studies establish the pertinence of pattern recognition to biological research and the applicability of pattern analysis in medical genomics and cancer research. Ordered nucleotide patterns emerging from seemingly random structures are prominent features of the interplay of nonlinear interactions among multiple heterogeneous components in DNA and RNA sequences, underlying biological processes such as duplication, segmentation, and unzipping [5]. Direct linear discriminant analysis (DLDA) enhances classification accuracy in DNA microarray gene expression data [6]. Transcription factor (TF) family-specific features by integrating DNA and protein-level information facilitate accurate classification and insight into TF-TFBS interactions [7]. Integration of Gene Ontology-based similarities within Bayesian networks to analyse protein-protein interaction networks outlines the role of probabilistic modelling in elucidating complex biological relationships [8]. A novel gene selection method using analytic hierarchy processes for microarray data classification is reported in [9]. Clustering can facilitate compression of sequences in increasingly growing biological databases [10], complement standard *alignment methods*, and shed light on the origin of physico-chemical interaction within genomic strands [11]. Key genes in breast cancer progression through genetic network analysis have been identified [12]. Patterns of distinctive nucleotide content in relation to local structure, composition, and mechanical strengths of DNA and RNA molecules are scrutinized [13]. Eukaryotic and prokaryotic genomes exhibit long range correlations quantified in terms of power law exponents of the probability distribution function of biological properties relevant to nucleosome positioning and mechanical properties as bending, torsional flexibility and propensity to form loops [14]. Shannon entropy of the basepair cluster distribution are reported for the 24 human chromosomes of the reference assembly GRCh37/hg19 [15]. The genomic distribution of protein coding segments and the chromosomal distribution of transposable elements (TEs) exhibit entropic scaling [16].

Entropy measures, self-similarity concepts, and scaling behaviour are well-established tools for characterizing intrinsic genomic features [17,18]. This work aims at investigating to which extent the human pangenome minigraphs preserve long-range correlation and self-similarity in comparison to the reference linear sequences. To that end, the probability distribution of the nucleotide segments yielded by

* Corresponding author.
E-mail addresses: renato.ferrero@polito.it (R. Ferrero), filippo.gandino@polito.it (F. Gandino), anna.carbone@polito.it (A. Carbone).

https://doi.org/10.1016/j.patrec.2025.03.004
Received 28 December 2023; Received in revised form 13 January 2025; Accepted 2 March 2025
Available online 12 March 2025
0167-8655/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

aligning the reference (GRCh38/hg38, T2T-CHM13) sequences and the 47 individuals assemblies, have been analysed for the 24 chromosomes. The nucleotide composition of the pangenome segments has been compared with the corresponding linear sequences. Quantifying the self-similarity of the minigraph segments is not a purely speculative issue as it can be adopted for optimizing segmentation, complementing traditional alignment methods, and accelerating the definition of the minigraph structure.

The remainder of the manuscript is organized as follows. Section 2 recalls the main steps of the pangenome construction and the numerical approach to transform the nucleotides text into a numerical sequence; Section 3 describes the information theoretical clustering approach; Section 4 illustrates the main outcomes of the work. Close-to-ideal power law scaling is obtained for the 24 chromosomes in the individual T2T-CHM13 reference sequences, whereas deviations from ideality are observed in the pangenome segments. Comments and future research directions can be found in Section 5.

2. Data

The Human Pangenome Reference Consortium (HPRC)[1] graph frameworks gather together segments extracted from genome sequences of 47 individuals [19]. Pangenomes are built by using different approaches: Minigraph, Minigraph-CACTUS and Pangenome Graph Builder (PGGB). The HPRC minigraphs are obtained by comparison to the reference GRCh38/hg38[2] and T2T-CHM13[3] genomic data. The procedure operates via an incremental algorithm based on genome-to-graph alignment to include the germline variants of the various individuals. An algorithm similar to *minimap2* is used to find local hits to segments in the graph. A *seed-and-extend* approach first finds common *seeds*, i.e., short segments common to the reference and the other genomic sequences, then extends and chains the seeds together in order to find longer and longer common segments. The resulting graph contains genome segments shared among different individuals. The segments obtained for the first 30 bases of chromosome 1 with the individual T2T-CHM13 taken as a reference are shown in Fig. 1.

The first step of the computational analysis carried out in this work is the numerical mapping of the reference sequences and of the pangenome segments, which are expressed in natural language texts, combination of letters corresponding to the nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). The standard representation of DNA and RNA bases by means of single characters was defined by the International Union of Pure and Applied Chemistry (IUPAC). The IUPAC nomenclature is a 16-character code representing single specifications for nucleotides (A, G, C, T/U) or ambiguity among 2, 3, or 4 possible nucleotides [21]. The IUPAC code is, in principle, case insensitive, but its established uses generally default to the capital case. A numerical value is assigned to either a single base (e.g., G for guanine, A for adenine) or a set of multiple bases (e.g. a couple R for either G or A and Y for either C or T or a triplet for codons etc.). Further generalizations of the IUPAC nomenclature have been proposed: one-dimensional, multidimensional, real or complex representations [22]. In this work, the nucleotides are mapped according to the RY rule: a puRine ($R = A, G$) is mapped to a positive value (+1), a pYrimidine ($Y = C, T$) is mapped to a negative value (−1) (Fig. 2). Next the set of +1 and −1 is summed and a random walk $y(x)$ (*DNA walk*) is obtained where x indicate the position of each nucleotide along the sequence. The RY rule is related to the DNA stiffness, which, in turn, depends on the types of base steps. In particular, the pyrimidine–purine

segment	*nucleotides*	*chromosome*	*offset*	*rank*
*s*1	*CTGAA*	*SN* : *Z* : *chr*1	*SO* :*i* : 0	*SR* :*i* : 0
*s*2	*ACG*	*SN* : *Z* : *chr*1	*SO* :*i* : 5	*SR* :*i* : 0
*s*3	*TGGC*	*SN* : *Z* : *chr*1	*SO* :*i* : 8	*SR* :*i* : 0
*s*4	*TGTGA*	*SN* : *Z* : *chr*1	*SO* :*i* : 12	*SR* :*i* : 0
*s*5	*TTTC*	*SN* : *Z* : *foo*	*SO* :*i* : 8	*SR* :*i* : 1
*s*6	*CTGA*	*SN* : *Z* : *foo*	*SO* :*i* : 12	*SR* :*i* : 1
*s*7	*GTTAC*	*SN* : *Z* : *bar*	*SO* :*i* : 5	*SR* :*i* : 2

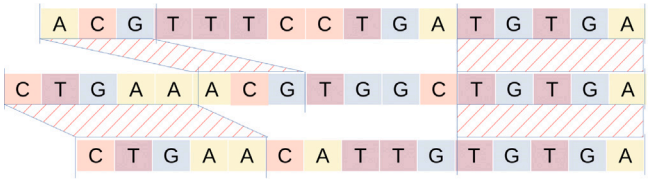


Fig. 1. First seven segments s1 to s7 of the minigraph corresponding to the first thirty bases of chromosome 1 [20]. The construction is based on the Graphical Fragment Assembly (GFA) reference standard. *S* identifies the segment, the other fields follow the TAG:TYPE:VALUE format. *SN* corresponds to the name of the stable sequence from which the segment is derived, *SO* to the offset on the stable sequence, *SR* to the rank (0 if on a linear reference genome), *Z* to string, *i* to signed integer. Nucleotide composition and alignment of segments s1 to s4 are shown at the bottom of the table.

Symbol	Basis	Meaning
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
R	A or G	puRine
Y	C or T/U	pYrimidine
M	A or C	aMino group
K	G or T/U	Keto group
S	C or G	Strong interaction (3 H bonds)
W	A or T/U	Weak interaction (2 H bonds)
H	*not* − *G*	A, C or T/U
B	not-A, B follows A	C, G or T/U
V	not T/U	A, C or G
D	not C	A, G or T/U
N	aNy	A, C, G or T/U

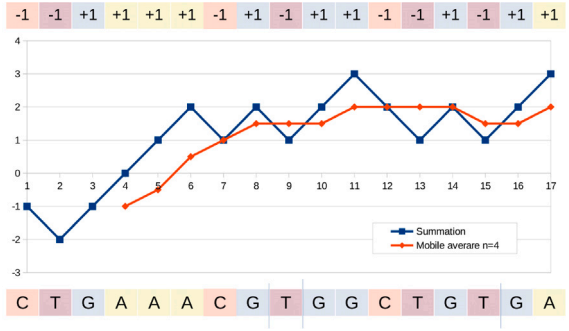


Fig. 2. (Top) IUPAC nomenclature with the 16-character alphabet. (Bottom) First bases of chromosome 1 mapped according to the RY rule: puRines (A, G) and pYrimidines (C, T) yield a sequence $R = +1$ and $Y = -1$ (first row on top) and then summed to yield the DNA walk (blue curve). The red curve shows a moving average $\widetilde{x}_i$ with $n = 4bp$.

(YR) and purine–pyrimidine (RY) base steps are respectively the least and the most thermally stable with highest persistence [23,24]. The RY rule has the advantage of helping to keep the nonstationarity of the numerical sequence at a minimum. The nonstationarity might be a serious drawback when long-range correlation, self-similarity and other statistical quantities are investigated. The average concentrations of A and T are about 0.30, those of G and C are about 0.20, hence the average concentrations of puRines (A,G) and pYrimidines (C,T)

[1] https://humanpangenome.org/

[2] https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph/hprc-v1.0-minigraph-grch38.gfa.gz

[3] https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph/hprc-v1.0-minigraph-chm13.gfa.gz

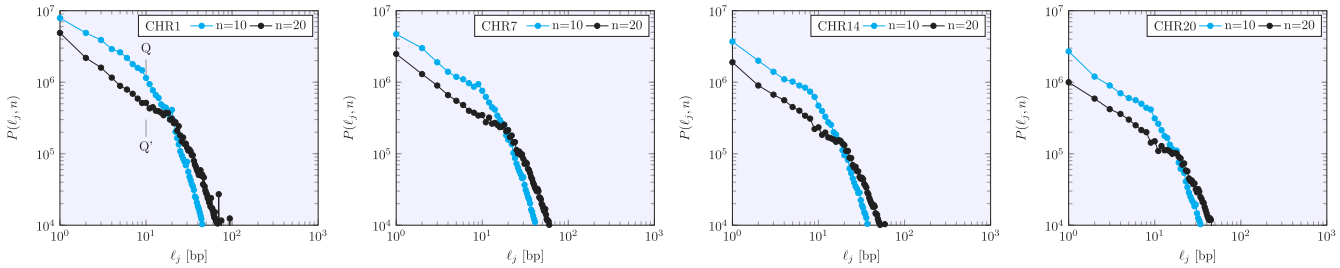


Fig. 3. Frequency of the cluster lengths $P(\ell_j, n)$ for chromosomes 1, 7, 14, 20 as defined in eq. (1) with $n = 10$ and $n = 20$. The probability distribution is not normalized.

are very close to 0.50 along the sequence. Therefore, coding puRines and pYrimidines to +1 / −1 yields a balanced numerical sequence. An unbalanced coding rule would amplify the strong variability of the local density distribution of the bases, giving rise to higher nonstationarity of the corresponding random walk. From the information-theoretical perspective, the extension of the nomenclature to map multiple nucleotides aims to assist and compress the representation of the rapidly growing space of genomic data. Hence, the identification of clusters/patterns of nucleotides gathering together information on meaningful DNA and RNA features is a very relevant interdisciplinary challenge.

3. Method

Statistical clustering, broadly referred to as partitioning a dataset into subsets according to some probabilistic criterion, is increasingly adopted to make sense of large amounts of data and to identify clusters naturally emerging from the data rather than by building artificial partitions [25–27]. Center-based clustering finds extensive applications, despite the pitfalls of suiting better to convex shaped clusters and requiring prior knowledge of the number of clusters. Density-based clustering identifies the clusters as high-density values of a given feature $\delta(x)$ surrounded by density values lower than a threshold $\bar{\delta}(x)$. Each intersection between the feature function $\delta(x)$ and the threshold $\bar{\delta}(x)$ generates separate connected regions in the feature domain. The threshold is usually taken as a *constant* over the whole feature domain, with the drawback that if its value is either too low or too high, several clusters will not show up. A constant threshold is an issue when the relevant feature is a long-range correlated quantity, a situation occurring in many real-world systems and in particular in biological contexts. Information-theoretic measures have been implemented to optimize the traditional clustering approaches giving rise to a fast developing area of computational science known as *information theoretic clustering* revolving around sample-by-sample estimates of the probability distribution function and information measures, such as Shannon, Kullback–Leibler (relative) and other entropy functionals [28].

The main steps and a few definitions of the *information theoretic clustering* approach [15,29,30] are briefly recalled hereafter. Consider a nucleotide sequence $\{x_i\}$ of length N and the local average $\widetilde{x}_{i,n} = \frac{1}{n}\sum_{n'=0}^{n-1} x(i-n')$ with $n \in (1, N)$. For each n, a partition $\{C\}$ of non-overlapping clusters is generated between consecutive intersections of $\{x_i\}$ and $\{\widetilde{x}_{i,n}\}$ defined by the nucleotide positions where the difference $\epsilon_{i,n} = x_i - \widetilde{x}_{i,n}$ is equal to zero. Hence, each cluster j is characterized by the random variable $\ell_j \equiv \|x_j - x_{j-1}\|$, with the instances x_{j-1} and x_j referring to subsequent intersection pairs. The random variable ℓ_j is the *cluster length*. The empirical distribution of the cluster lengths $P(\ell_j, n)$ is obtained by ranking the clusters $\mathcal{N}(\ell_1, n), \mathcal{N}(\ell_2, n), \ldots, \mathcal{N}(\ell_j, n)$ according to their length for each n:

$$P(\ell_j, n) = \frac{\mathcal{N}(\ell_j, n)}{\mathcal{N}_C(n)} \tag{1}$$

with $\mathcal{N}_C(n) = \sum_{j=1}^{k(n)} \mathcal{N}(\ell_j, n)$ the number of clusters generated by the partition for each n. Let $k = \sum_{n=1}^{N} \mathcal{N}_C(n)$ indicate the total number of clusters for all the possible values of n and the normalization condition holds as usual:

$$\sum_{n=1}^{N}\sum_{j=1}^{\mathcal{N}_C(n)} P(\ell_j, n) = 1 \quad , \tag{2}$$

where the index j runs over the clusters obtained by each partition with size n with $n \in (1, N)$. By introducing $P(\ell_j, n)$ in the Shannon functional, the *cluster entropy* writes:

$$S_C[P] = -\sum_{j,n} P(\ell_j, n) \log P(\ell_j, n) \quad . \tag{3}$$

The meaning of the cluster entropy can be discussed by approximating the Shannon entropy in Eq. (3) in the limit of continuous variables, yielding the *differential cluster entropy*. The probability distribution function $P(\ell, n)$ of the lengths ℓ for each n can be written as:

$$P(\ell, n) \sim \ell^{-D} \mathcal{F}(\ell, n) \sim \mu(\ell, n)^{-1} \quad . \tag{4}$$

The function $\mathcal{F}(\ell, n)$ in Eq. (4) accounts for the finiteness of the cluster size and can be taken of the form: $\mathcal{F}(\ell, n) \sim \exp(-\ell/n)$. For long-range correlated sequence described as Fractional Brownian Motion, the exponent $D = 2 - H$ is the fractal dimension of the sequence with H the Hurst exponent. The exponents D and H are widely used for quantifying long-range correlations (power-law decaying distribution) in complex systems. The differential entropy writes:

$$S(\ell, n) \sim S_0 + \log \ell^D + \frac{\ell}{n} \quad , \tag{5}$$

where S_0 is a constant, $\log \ell^D$ is related to the term ℓ^{-D} and ℓ/n is related to the term $\mathcal{F}(\ell, n)$.

According to Eq. (4), $P(\ell, n)$ exhibits two regimes characterized by different slopes [15]. At short cluster length ($\ell < n$), the probability distribution function is power-law sloped. For ℓ larger than n, the probability distribution function decreases more quickly, approaching an exponential behaviour related to the finite-size effects. The power-law behaviour of the probability distribution $P(\ell, n)$ determines the cluster entropy $S(\ell, n)$ to increase as $\log \ell^D$ and be n-invariant for small values of ℓ. Conversely, $S(\ell, n)$ increases as ℓ/n, i.e. a linear function of ℓ, at larger ℓ, as expected according to Eq. (5). Clusters with lengths ℓ larger than n are not power-law correlated, due to finite-size effects. Hence, they are characterized by entropy values exceeding the curve $\log \ell^D$, which corresponds to power-law correlated clusters. The power-law scaling of the cluster length distribution $P(\ell, n)$ is reflected in the logarithmic behaviour of the cluster entropy $S(\ell, n)$ with ℓ the length of the clusters and n the moving average window. These features are related to long-range order and fractality. The transition from compositionally ordered to disordered cluster has been visualized for the 24 chromosomes. The occurrence of long-range correlations implies that the nucleotides self-organize and give rise to a *compositional self-similarity* along the chromosome sequences [15].

The empirical probability distribution $P(s_j)$ of the pangenome segments with lengths $s_1, s_2, s_3, \ldots, s_j$ extracted from 47 individuals as given in [20] is estimated by counting the number of segments

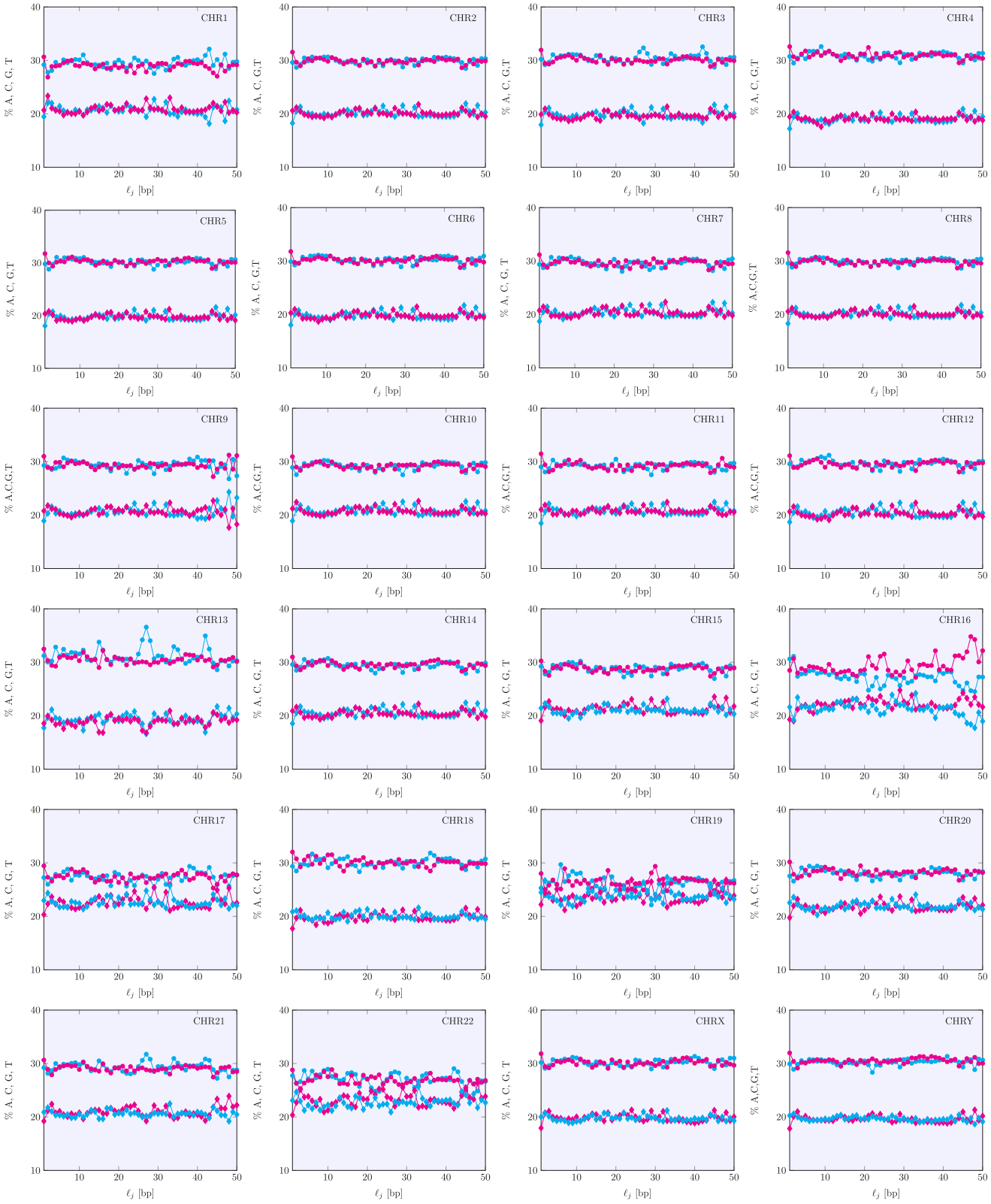


Fig. 4. Percentage of puRines ($R = A, G$) and pYrimidines ($Y = C, T$) versus cluster lengths ℓ_j for the 24 chromosomes of the linear sequences of the T2T-CHM13 individual with $n = 20$. Symbols refer respectively to nucleotides A ○, T ●, C ◆, G ◇.

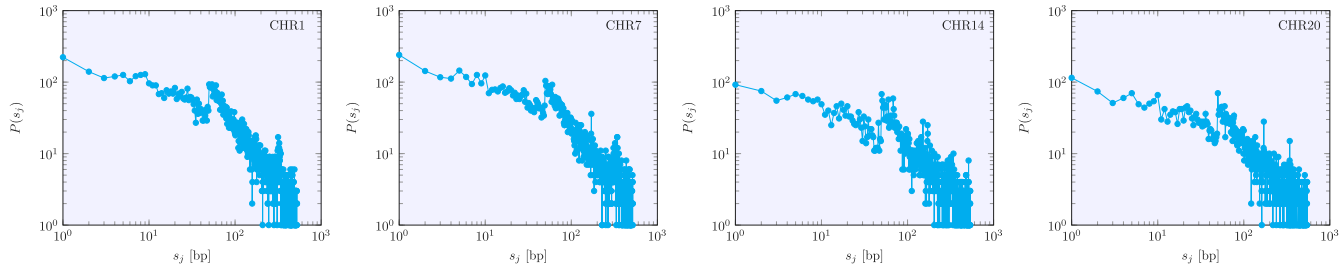


Fig. 5. Frequency count of the pangenome segments $P(s_j)$ for the minigraphs of chromosomes 1, 7, 14, 20, built with reference to the individual genome T2T-CHM13. Frequency count is not normalized.

$s_1, s_2, s_3, \ldots, s_j$ collected in the minigraph of each chromosome according to their length:

$$P(s_j) = \frac{\mathcal{N}(s_j)}{\mathcal{N}_S} \tag{6}$$

where $\mathcal{N}(s_j)$ indicates the number of segments with length s_j and $\mathcal{N}_S$ indicates the total number of segments, with any length, corresponding to each of the 24 chromosomes in the pangenome. Samples of the segments s_j for the first bases of chromosome 1 are shown in the table of Fig. 1. The alignment between the T2T-CHM13 reference sequence (centre) and other two sequences (above and below) is illustrated at the bottom of the table. As a final note, we observe that the parameter n does not show up in $P(s_j)$ (Eq. (6)) contrarily to the probability distribution function $P(\ell_j, n)$ defined in Eq. (1). The reason is that $P(s_j)$ counts the segments defined by the pangenome minigraph construction without using the moving average window n.

In the following Section, the information theoretic method based on the Shannon entropy will be implemented on the 24 chromosomes of the pangenome minigraphs and of the reference individuals (GRCh38/hg38 and T2T-CHM13). The analysis of the GRCh38/hg38 and T2T-CHM13 linear sequences is needed to assess the degree of long-range correlation and how this is reflected in the compositional self-similarity of the clusters in terms of percentage of the nucleotides A, C, G, T. The main purpose is the assessment of the self-similarity and power-law scaling of the reference individuals and the establishment of a benchmark for the pangenome segments.

4. Results

As above mentioned, the long-range correlation analysis is performed on the pangenome minigraph segments by taking the reference genomes of the GRCh38/hg38 and T2T-CHM13 individuals as benchmarks. To this purpose, dedicated Python codes have been implemented with the ability to extract and analyse the data from the repositories[4]. The probability distribution function $P(\ell_j, n)$, defined in Eq. (1), has been estimated by counting the clusters generated in the linear sequences of the 24 chromosomes of the reference individuals according to the method discussed in Section 3. The results referred to the sequences of chromosomes 1, 7, 14, 20 of the T2T-CHM13 individual are shown in Fig. 3. The curves are the probability distribution of the clusters obtained by using the moving average partition for $n = 10$ and $n = 20$. It is worthy to note that clusters with the same length can be generated by different values of the parameter n. For example, consider the probability distribution function of chromosome 1 shown in Fig. 3, clusters with $\ell_j = 10bp$ have pdf corresponding either to the point Q (with $n = 10bp$) or Q' (with $n = 20bp$). However, Q' corresponds to power-law correlated (ordered) clusters, since Q' lays on the curve that scales as a power law, according to the description using continuous variables ℓ^D, ($\ell = 10bp < n = 20bp$). In contrast, Q lies on the exponentially decaying portion of the pdf $F(\ell, n)$ which originates from $\ell_j > n = 10bp$. In summary, clusters with lengths shorter than n correspond to ordered (long-range correlated) nucleotides, whereas clusters with lengths larger than n correspond to disordered (exponentially correlated) nucleotides. The behaviour of the probability distribution functions shown in Fig. 3 is reflected in the compositional features of the clusters. Power-law correlation implies self-similarity of the cluster sequence structure not only at the numerical and geometrical level but also in terms of chemical and structural properties. The nucleotide composition in terms of percentage of A, C, T, G is shown in Fig. 4 for all 24 chromosomes. One can note that self-similarity holds over a broad range of lengths ℓ_j for clusters generated with $n = 20bp$.

Next, we analyse the long-range correlation properties of the pangenome minigraph segments. Fig. 5 shows the frequency $P(s_j)$, defined by Eq. (6), respectively for chromosomes 1, 7, 14 and 20 extracted by the human pangenome minigraph. The values of the frequency $P(s_j)$ have been obtained by counting the segments generated in the chromosome minigraphs according to their length. The first bases and the first four segments are shown in Fig. 1 for the chromosome 1 minigraph. The frequency $P(s_j)$ exhibits a discontinuity with a slope change at segments lengths on the order of $50bp$, a fact that can be related to the different segmentation methods respectively adopted for segments shorter/longer than $50bp$ as described in [1]. Artificial partitions unavoidably exhibit discontinuity reflected in the onset of two scaling regimes. The nucleotide percentage in the segments with length $s_j < 50bp$ is plotted in Fig. 6 for all the 24 chromosomes. One can note a large variability in the nucleotide composition in comparison to what has been found for the linear sequences of the reference chromosomes shown in Fig. 4. An analogous level of variability is found for the other chromosomes as can be seen in the corresponding figures. The compositional variability of the minigraph segments observed in this work, taking as reference the T2T-CHM13, is also larger than the variances of the GHRc37/h19 24 chromosomes, estimated in [15].

5. Conclusions

A systematic analysis of long-range correlation and compositional self-similarity of the minigraphs corresponding to the 24 chromosomes of the recently published human pangenome has been performed. The probability distribution $P(s_j)$ of the segment lengths s_j of the human pangenome has been estimated to assess correlation and self-similarity. As a benchmark, the probability distribution $P(\ell_j, n)$ of the clusters with length ℓ_j has been evaluated for the reference linear sequences of the individuals GRCh38/hg38 and T2T-CHM13. The power-law scaling properties and the nucleotide composition of the pangenome segments have been compared to those of the long-range correlated clusters obtained from the linear sequence of reference.

The aim of the work was to assess to what extent long-range correlations are exhibited by the segments composing the minigraphs of the human pangenome, and to find out whether the segmentation process maintains the character of self-similarity observed in the corresponding linear sequences. Our work reveals that the segmentation process does

[4] Python code and further results for other reference chromosomes can be found at https://github.com/rntf/pangenome-clustering



Fig. 6. Percentage of puRines ($R = A, G$) and pYrimidines ($Y = C, T$) versus segment lengths s_j for the 24 chromosomes of the pangenome minigraph. The segments are obtained by aligning the chromosomes of the 47 individuals to those of the linear sequences of the T2T-CHM13 individual. Symbols refer to the nucleotides A ●, T ●, C ◆, G ◆.

not fully preserve the self-similarity of the segments as shown by the limited extent of the power law decay of the probability distribution function and the onset of two range of decays at short and long fragment lengths. By reverting the numerical sequence of each segment to the nucleotide alphabet, it has been observed that the nucleotide composition still exhibits self-similar features though to a lower extent compared to the continuously segmented clustering. This implies that a significant amount of genomic information or gene content is variable among closely related individuals of a species and might help address some open questions regarding the pangenome building principles.

The number of individuals needed for a comprehensive pangenome of a given species is currently debated, as it depends on the diversity and resolution of genetic variation one aims to capture. Defining a set of quantitative indexes derived from the scaling behaviour of the relevant probability distributions can provide a theoretical estimation of this amount. Additionally, the scaling behaviour derived from probability distributions offers a promising tool for diagnosing genomic variability and guiding decisions about the inclusion of additional genomes. The main open question concerns the selection of criteria for segmenting and constructing the pangenome structure. The analysis can be viewed as a tool to quantify the intrinsic similarity between the individual genomes constituting the minigraph. The geometrical self-similarity of the minigraph segments can yield the compositional self-similarity of the segments in terms of nucleotide percentage. These findings suggest that the segment self-similarity, quantified through power-law correlations and related exponents, could serve as a foundational criterion for segmenting and constructing the pangenome. This approach might have meaningful implications from a biological perspective: regions of the genomes characterized by a high level of heterogeneity and disorder are more likely to duplicate and mutate.

Further developments of the proposed approach have been envisioned. The generalization of the information theoretic clustering approach would allow for more robust statistical inference about the pangenome structure and included individuals. Estimates of relative cluster entropy [29] could be used to quantify the distance between the empirical probability distribution of the minigraph segments $P(s_j)$ and the cluster distribution $P(\ell_j, n)$ of the reference linear sequences. The minimum relative entropy principle could be adopted as a more stringent optimization criterion for statistical inference about pangenome structure. This study demonstrates the value of using self-similarity and scaling laws in tandem with information theoretic tools as quantitative metrics for pangenome analysis while emphasizing the need for refined tools and models to address remaining challenges.

CRediT authorship contribution statement

Renato Ferrero: Software, Methodology, Data curation. **Filippo Gandino:** Software, Resources, Data curation. **Anna Carbone:** Writing original draft, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Anna Carbone reports financial support was provided by European Commission. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge support from the TED4LAT project, a WIDERA initiative within the Horizon Europe Programme, Grant Agreement 101079206.

Data availability

Data used in this work (Human Pangenome, GRCh38 and CHM13 chromosome sequences are open access. The urls of the data repositories are provided in the manuscript.

References

[1] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J.K. Lucas, J. Monlong, H.J. Abel, et al., A draft human pangenome reference, Nature 617 (7960) (2023) 312–324.
[2] T.L. Pedersen, Hierarchical sets: analyzing pangenome structure through scalable set visualizations, Bioinformatics 33 (11) (2017) 1604–1612.
[3] V. Bonnici, R. Giugno, V. Manca, PanDelos: a dictionary-based method for pangenome content discovery, BMC Bioinformatics 19 (15) (2018) 47–59.
[4] C.A. Page, A.J. Cummins, M. Hunt, V.K. Wong, S. Reuter, M. T.G. Holden, M. Fookes, D. Falush, J.A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pangenome analysis, Bioinformatics 31 (22) (2015) 3691–3693.
[5] T. Misteli, The self-organizing genome: principles of genome architecture and function, Cell 183 (1) (2020) 28–45.
[6] K.K. Paliwal, A. Sharma, Improved direct lda and its application to dna microarray gene expression data, Pattern Recognit. Lett. 31 (16) (2010) 2489–2492.
[7] G.B. Fogel, A. Anand, G. Pugalenthi, P.N. Suganthan, Identification and analysis of transcription factor family-specific features derived from dna and protein information, Pattern Recognit. Lett. 31 (14) (2010) 2097–2102.
[8] H. Wang, H. Zheng, F. Browne, D.H. Glass, F. Azuaje, Integration of gene ontology-based similarities for supporting analysis of protein–protein interaction networks, Pattern Recognit. Lett. 31 (14) (2010) 2073–2082.
[9] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, A novel aggregate gene selection method for microarray data classification, Pattern Recognit. Lett. 60 (2015) 16–23.
[10] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (3) (2001) 282–283.
[11] K.S. Bohnsack, M. Kaden, J. Abel, T. Villmann, Alignment-free sequence comparison: A systematic survey from a machine learning perspective, IEEE/ ACM Trans. Comput. Biology Bioinform. (2022).
[12] I.U. Martínez Vargas, M.O. León Pineda, M. Alvarado Mentado, Main genes in breast cancer primary tumor and first metastasis in lymph nodes revealed by information-theory-based genetic networks pattern analysis, Pattern Recognit. Lett. 186 (2024) 369–376.
[13] P. Bernaola-Galván, P. Carpena, C. Gómez-Martín, J.L. Oliver, Compositional Structure of the Genome: A Review, Biology 12 (6) (2023) 849.
[14] M.O. Costa, R. Silva, D.H.A.L. Anselmo, J.R.P. Silva, Analysis of human DNA through power-law statistics, Phys. Rev. E 99 (2) (2019) 022112.
[15] A. Carbone, Information measure for long-range correlated sequences: the case of the 24 human chromosomes, Sci. Rep. 3 (1) (2013) 2721.
[16] D. Polychronopoulos, L. Athanasopoulou, Y. Almirantis, Fractality and entropic scaling in the chromosomal distribution of conserved noncoding elements in the human genome, Gene 584 (2) (2016) 148–160.
[17] X. Lin, Y. Qi, A.P. Latham, B. Zhang, Multiscale modeling of genome organization with maximum entropy optimization, J. Chem. Phys. 155 (1) (2021).
[18] S. Kak, Self-similarity and the maximum entropy principle in the genetic code, Theory Biosci. 142 (3) (2023) 205–210.
[19] H. Li, X. Feng, C. Chu, The design and construction of reference pangenome graphs with minigraph, Genome Biology 21 (2020) 1–19.
[20] G. Hickey, J. Monlong, J. Ebler, A.M. Novak, J.M. Eizenga, Y. Gao, T. Marschall, H. Li, B. Paten, Pangenome graph construction from genome alignments with Minigraph-Cactus, Nature Biotechnol. (2023) 1–11.
[21] A. Cornish-Bowden, Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984, Nucleic Acids Res. 13 (9) (1985) 3021.
[22] A.D. Johnson, An extended IUPAC nomenclature code for polymorphic nucleic acids, Bioinformatics 26 (10) (2010) 1386–1389.
[23] S.A. Ishikawa, Y. Inagaki, T. Hashimoto, RY-coding and non-homogeneous models can ameliorate the maximum-likelihood inferences from nucleotide sequence data with parallel compositional heterogeneity, Evol. Bioinform. (2012) 8:EBO–S9017.
[24] D. Cohen, General designs reveal distinct codes in protein-coding and non-coding human DNA, Genes 13 (11) (2022) 1970.
[25] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.
[26] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, J. Mach. Learn. Res. 11 (2010) 2837–2854.
[27] J. Bailey, M.E. Houle, X. Ma, Relationships between tail entropies and local intrinsic dimensionality and their use for estimation and feature representation, Inf. Syst. 118 (2023) 102245.
[28] E. Gokcay, J.C. Principe, Information theoretic clustering, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 158–171.
[29] A. Carbone, L. Ponta, Relative cluster entropy for power-law correlated sequences, SciPost Phys. 13 (3) (2022) 076.
[30] A. Carbone, G. Castelli, H.E. Stanley, Analysis of clusters formed by the moving average of a long-range correlated time series, Phys. Rev. E 69 (2) (2004) 026105.