http://dx.doi.org/10.21622/ACE.2021.01.1.007

Computing Genomic Signatures Using de Bruijn Chains

Lenwood S. Heath[1*], Amrita Pati[2]

[1] Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2] Amgen Corporation, California, USA
* heath@vt.edu

Email: {heath@vt.edu, apati@amgen.com}

Received on, 29 March 2021 - Accepted on, 02 May 2021 - Published on, 16 June 2021

Abstract

Genomic DNA sequences have both deterministic and random aspects and exhibit features at numerous scales, from codons to regions of conserved or divergent gene order. Genomic signatures work by capturing one or more such features efficiently into a compact mathematical structure. We examine the unique manner in which oligonucleotides constitute a genome, within a graph-theoretic setting. A de Bruijn chain (DBC) is a kind of de Bruijn graph that includes a finite Markov chain. By representing a DNA sequence as a walk over a DBC and retaining specific information at nodes and edges, we obtain the de Bruijn chain genomic signature θ^{dbc}, based on graph structure and the stationary distribution of the DBC. We demonstrate that the θ^{dbc} signature is information-rich, efficient, sufficiently representative of the sequence from which it is derived, and superior to existing genomic signatures such as the dinucleotide odds ratio and word frequency-based signatures. We develop a mathematical framework to elucidate the power of the θ^{dbc} signature to distinguish between sequences hypothesized to be generated by DBCs of distinct parameters. We study the effect of order of the θ^{dbc} signature, genome size, and variation within a genome on accuracy. We illustrate its superior performance over existing genomic signatures in predicting the origin of short DNA sequences.

Introduction

The genome $\mathcal{G}$ of an organism is a set of long nucleotide sequences modeled, within a formal language framework, as strings over $\Sigma_{\text{DNA}} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$, the DNA alphabet. Every genome has a unique constitution of nucleotides that encode specific phenotypic traits and regulate the cellular and biological processes of that organism. Unique features of a genomic sequence that are globally conserved and can be captured in the form of mathematical structures can serve as signatures for that genome. Since $\mathcal{G}$ itself differs from one organism to another, it can serve as a unique mathematical structure representing an organism. However, a genome is typically quite large (e.g., billions of bases for the human genome) and also demonstrates slight differences from one individual of a species to another. Fix a genomic sequence H that is a substring of some string in $\mathcal{G}$. Intuitively, a *genomic signature* for an organism is a mathematical structure $\theta(H)$ derived from H, which, ideally, can be efficiently computed, is significantly smaller to represent than H, and, if H is sufficiently representative of $\mathcal{G}$, can accurately identify the original organism. The intent is that the signature of other large substrings from $\mathcal{G}$ be highly similar to $\theta(H)$ and distinguishable from signatures of other organisms. A genomic signature is judged along two, typically antagonistic, dimensions: (1) the amount of compression achieved by $\theta(H)$, and (2) its effectiveness in identifying the genome.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

The term "genomic signature" must not be confused with the term "gene expression signature" *[1, 2]* although the two terms have been used interchangeably in a few works [3–7]. A *gene expression signature* is a distinct conserved model of gene expression patterns observed in a set of genes during specific biological phenomena or environmental conditions [1, 2]. Normark et al. [8] have used the term "genomic signature" to represent long term genomic effects of the loss of sex and recombination on asexual eukaryotic genomes. Cannon et al. [9] have used it to represent probe sequences that are short (25 bp and less) primers that are hyperdispersed in a probability space of sequences and generated without the knowledge of the target genome, while scientists who study the effects of ionizing radiation on genomes use the term to indicate radiation-induced genomic changes such as gene copy number and intrachromosomal aberrations [10, 11]. For the purposes of this paper, a genomic signature, as defined in the previous paragraph, is a unique mathematical structure strictly computed from sequence data and conserved for reasonably large ($\geq$ few kilobases) subsequences of a genome for a wide range of subsequence lengths.

In this paper, we propose a novel genomic signature called the de Bruijn chain signature θ^{dbc}. A de Bruin chain (DBC) is a de Bruijn graph with an underlying finite Markov Chain. We derive the θ^{dbc} signature by thinking of a genomic sequence at hand as a walk over a suitably defined DBC. We then include characteristic properties of the stationary distribution of the underlying Markov chain and the manner in which the DBC disintegrates on deleting edges by a systematic method, as components of the θ^{dbc} signature. By definition, the θ^{dbc} signature retains features of genomic sequences that are different from features retained by word-count based signatures explored in related literature. Here, we explore the properties of the θ^{dbc} signature and several other genomic signatures with an emphasis on the identification of short unknown DNA sequences.

The species from which a genomic sequence is derived is its *origin*. A genomic sequence X of unknown origin is to be analyzed. We visualize X as an overlap of numerous successive short sequences of pre-defined length w each, in a specific manner. The *order* is the above word length w at which a genomic sequence is analyzed. A pre-defined signature $\theta_w(X)$, at order w, is computed from X and compared to the same signature at the same order w for all available species. The correlations between $\theta(X)$ and the existing signatures are used to predict the origin of X. We demonstrate that the θ^{dbc} signature performs better than its competitors, the dinucleotide odds ratio θ^{dor} and the word count vector θ^{wcv}. We further illustrate that combining the strengths of the θ^{dbc} signature and the θ^{dor} signature results in higher accuracy of origin identification while distinguishing between distant species.

Some applications of genomic signatures are as follows. A database of signatures of all fully or partially sequenced genomes can be constructed. Apart from being a beneficial public resource, such a database will enable identification of the origin and/or closest relatives of segments of unknown DNA. An exhaustive database will lead to the discovery of new species and their placement on the tree of life [12]. A sequence identification gadget constructed using this database and the algorithms we propose can be used as a household utility for testing food products for infectious microbial growth, screening insects for parasites, and understanding the origin and properties of plants and animals in the surroundings. Such an instrument will be invaluable to ecologists. It is also possible to apply genomic signatures to binning metagenomic data.

This paper is organized as follows. The Related Work Section reviews the relevant literature. In the Preliminaries Section, we define relevant mathematical concepts and establish notation. We also introduce graph-based signatures and define the DBC signature θ^{dbc} in this section. In the Results Section, we present evidence of the efficiency of the θ^{dbc} signature in identifying origins of short sequences and illustrate its superior performance over existing signatures. In the Theory Section, we derive theoretical bounds that characterize the abilities of the word count vector signature θ^{wcv} and the θ^{dbc} signature to differentiate between genomes. Conclusions and future directions are presented in the Conclusions Section.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

Related Work

A *DNA word* or an *oligonucleotide* is a short string of predefined order over the DNA alphabet. Oligonucleotide frequencies have been described as characteristic features of genomes in many works [13–25]. Karlin and Burge [19] were among the first to use the term *genomic signature*. They define the *dinucleotide odds ratio* (θ^{dor}) or *relative abundance*, which is the collection of 16 functions defined for dinucleotides XY by $\rho_{XY}(H) = \frac{\text{freq}(XY, H)}{\text{freq}(X, H)\,\text{freq}(Y, H)}$ where $\text{freq}(x, H)$ is the frequency of string x as a substring in H. They observe that ρ values are similar throughout a genome and compare θ^{dor} for a few organisms to demonstrate its capability of distinguishing organisms. Karlin et al. [20] observe that individual components of the θ^{dor} vector typically range from 0.78 to 1.23. They use a normalized L_1-distance, called *delta-distance* (δ), to distinguish between species. The δ-distance between the θ^{dor} signatures of sequences H_1 and H_2 is defined as $\delta(H_1, H_2) = \frac{1000}{16} \sum_{XY \in \mathcal{S}^2} |\rho_{XY}(H_1) - \rho_{XY}(H_2)|$. Campbell et al. [26] compare θ^{dor} signatures of prokaryotic, plasmid, and mitochondrial DNA. Gentles and Karlin [27] examine the θ^{dor} signature in sequences of eukaryotic genomes and chromosomes, including human chromosomes 21 and 22, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Drosophila melanogaster*. Jernigan and Baran [18] demonstrate empirically that the δ-distance between θ^{dor} signatures of strings sampled within a genome is approximately preserved over a wide range of string lengths, while it varies for strings sampled from different genomes.

The second most widely-used method in the literature to visualize and study the composition of and separation between DNA sequences is the Chaos Game Representation (CGR) signature [28–30]. Mathematically, the Chaos Game is an iterated function system that uses a two-dimensional heat-map-style plot to provide a visual representation of composition of a given DNA sequence through tiled geometrical patterns that sharpen with increasing DNA word lengths. Deschavanne et al. [15] constructed CGR images from oligonucleotide frequencies and built the application GENSTYLE [15, 17], which predicts the approximate origin of a sequence using L_1-distances to oligonucleotide frequency vectors of all genome sequences in the Entrez database, thus formally introducing CGRs as a genomic signature.

The simple word count vector has also been used as a genomic signature in various works. The application TETRA (Teeling et al. [22]) uses tetranucleotide frequencies to calculate similarity between sequences. For bacterial species, Coenye and Vandamme [14] correlate δ with 16S rDNA sequence similarity and DNA-DNA hybridization values. For 57 prokaryotic genomes, Sandberg et al. [21] compare G+C content, oligonucleotide frequency, and codon bias. Dufraigne et al. [16] and van Passel et al. [23] employ oligonucleotide frequencies to identify regions of horizontal gene transfer (HGT) in prokaryotes. Carbone et al. [13] correlate the ecological niches of 80 Eubacteria and 16 Archaea to codon bias used as a genomic signature.

The genomic signatures outlined above demonstrate that signatures differ among species, but, with the exception of the dinucleotide odds ratio, no one else formally addresses the amount of variation, identification of unknown DNA, and the effect of short available sequence length on these signatures. As part of our DNA Words program investigating mathematical invariants derived from genomes, we examine the finest scale in graph-theoretic terms, while integrating DNA word graph structure with Markov chain properties. One frequently exploited observation is that a string over Σ_{DNA} defines a walk in a suitably defined de Bruijn graph. Closely related is the correspondence of such a string to an Eulerian tour in a suitably defined multigraph. Applications include DNA physical mapping, DNA sequence assembly, and multiple sequence alignment problems [31–35]. In Heath and Pati [36], we explore purely graph-based genomic signatures and compare their performance with the word count vector and the dinucleotides odds ratio signatures. We identify a graph-based signature that is competitive with the dinucleotides odds ratio (most efficient among existing signatures), performing marginally better. Subsequently [37], we introduce the de Bruijn

http://dx.doi.org/10.21622/ACE.2021.01.1.007

chain signature θ^{dbc} and demonstrate that it performs better than all existing genomic signatures with emphasis on target identification from short DNA segments. This signature performs much better than oligonucleotide frequency vectors in differentiating among diverse genomes. In this work, we propose a mathematical framework for characterizing the ability of the θ^{dbc} signature to distinguish between genomes using short genomic segments. We examine the effect of different orders on the efficiency of the θ^{dbc} signature. We also study relationships among efficiency, genome variation, and genome size. Also, see our subsequent work [38, 39] and Pati's dissertation [40].

There has been little advancement in the concept of genomic signatures since our work, though there has continued to be investigations into comparing genomic sequences for similarity and dissimilarity. Konstantinidis and Tiedje [41] and Goris et al. [42] proposed to measure the similarity of two genomes through the average amino acid content (AAI) and the average nucleotide content (ANI); ANI is measured via a heuristic that employs either BLAST [43] or MUMmer [44]. Pritchard, et al. [45] provide a flexible implementation of the ANI heuristic through the Python tool pyani. Initially using ANI as the basis of genome comparisons, Vinatzer and Heath [46–51] have developed the Life Identification Number (LIN) concept that provides a framework for identifying and naming all sequenced genomes. Broder [52] introduced the MinHash concept for measuring the similarity of documents; it provides an estimate of the Jaccard similarity between the word contents of two documents. Ondov et al. [53] adapted MinHash to estimate a distance between two genomes using the set of k-mers present in the two genomes. The k-mer and MinHash concepts have been further developed by numerous researchers since then [54–57].

Preliminaries

An *alphabet* Σ is a finite, non-empty set of symbols. The *binary alphabet* is $\Sigma_{\mathsf{B}} = \{0, 1\}$, while the *DNA alphabet* is $\Sigma_{\mathsf{DNA}} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. A *string* or *word* x over Σ is a finite sequence $x = \sigma_1\sigma_2 \cdots \sigma_w$ of symbols from Σ; its *length* $|x|$ is w. A single chromosome in a genome is typically written as the string over Σ of nucleotides on one DNA strand. A *genomic sequence* is a chromosomal sequence or any substring of it. An organism's genome $\mathcal{G}$ is the set of all its chromosomal sequences.

For strings x and y, $\mathsf{occ}(x, y)$ is the count of occurrences of x as a substring of y. If $|x| \leq |y|$, the *frequency* of x in y is $\mathsf{freq}(x, y) = \dfrac{\mathsf{occ}(x, y)}{|y| - |x| + 1}$. Fix a word length $w \geq 1$. Let $l = 4^w$. The *order-w state space* is $\mathcal{S}^w = \Sigma^w_{\mathsf{DNA}}$, the set consisting of the l words of length w. For $1 \leq i \leq l$, let x_i be the i^{th} element of $\mathcal{S}^w$ in lexicographic order.

The *order-w de Bruijn graph* $\mathcal{DB}^w = (\mathcal{S}^w, E)$ over alphabet Σ is a directed graph, where $(x_i, x_j) \in E$ when $x_i\sigma = \iota x_j$, for some $\sigma, \iota \in \Sigma$; such an edge is labeled σ (see [58]). Figure 1 depicts the de Bruijn graphs of order 3 over the binary alphabet $\Sigma_{\mathsf{B}} = \{0, 1\}$. As observed, the vertex set of the binary de Bruijn graphs of order 3 is the set $(\{000, 001, 010, 011, 100, 101, 110, 111\})$ of all binary strings of length 3. The vertex set of the DNA de Bruijn graph of order 2 is $\{\mathsf{AA}, \mathsf{AC}, \mathsf{AG}, \mathsf{AT}, \mathsf{CA}, \mathsf{CC}, \mathsf{CG}, \mathsf{CT}, \mathsf{GA}, \mathsf{GC}, \mathsf{GG}, \mathsf{GT}, \mathsf{TA}, \mathsf{TC}, \mathsf{TG}, \mathsf{TT}\}$.

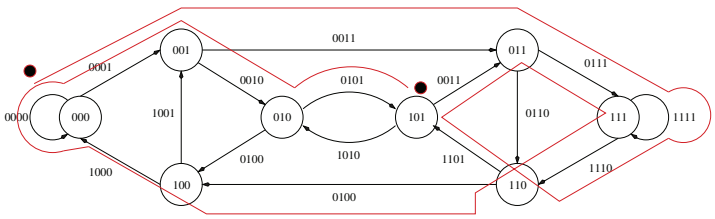


Fig 1. The order-3 de Bruijn graph on the binary alphabet; the red line indicates a walk in the graph traced by the sequence 0001110111000101.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

Let $H \in \Sigma^*_{\text{DNA}}$ have length $|H| = n$; we think of H as a long genomic sequence that traces a walk in $\mathcal{DB}^w$. The *vertex count* of x_i in H is $\text{vc}(x_i, H) = \text{occ}(x_i, H)$, while the *edge count* of edge $(x_i, x_j) \in E$ in H, where $x_i\sigma = \gamma x_j$, is $\text{ec}((x_i, x_j), H) = \text{occ}(x_i\sigma, H)$. The *order-$w$ word count vector* $\theta^{wcv}_w(H)$ of H is the l-vector having components $\text{occ}(x_i, H)$, in lexicographic order. The corresponding *order-w word frequency vector* is the l-vector having components $\text{freq}(x_i, H)$, in lexicographic order. In Figure 1(b), for instance, the word count vector is $\langle 2, 2, 1, 2, 1, 2, 2, 2\rangle$. Nucleotide frequencies vary among organisms, while, as Fickett et al. [59] observe, the frequencies of A's and T's (and hence of G's and C's) are approximately constant within a single genome.

Now consider the Markov chain underlying the above de Bruijn graph $\mathcal{DB}^w = (\mathcal{S}^w, E)$. That Markov chain has state space $\mathcal{S}^w$ and a sparse transition probability matrix with nonzero transition probabilities only for edges in $\mathcal{DB}^w$; such a Markov chain is called an *order-w de Bruijn chain (DBC)*. In this paper, we use DBCs in modeling of genomic signatures, based on the following intuition. Let $\mathcal{DC}$ be an order-w DBC with $l \times l$ transition probability matrix $P = (p_{ij})$; here, p_{ij} is the probability of a one-step transition from state x_i to state x_j [60]. P is sparse, with at most 4 nonzero entries per row. The *order-w DBC, $\mathcal{DC}^w(H)$, for genomic sequence H* has transition probabilities

$p_{ij} = \dfrac{\text{ec}((x_i, x_j), H)}{\text{occ}(x_i, H)}$, $\text{occ}(x_i, H) > 0$. Genomic sequences are sufficiently large and diverse in their composition to ensure occurrence of all words in $\mathcal{S}^w$ for reasonably small $w \in [1..5]$. Any DBC generating such a sequence is irreducible. We also assume that DBCs generating genomic sequences are aperiodic with finite state space, and hence, recurrent non-null. Thus, we assume that all DBCs are ergodic and hence that there is a unique *stationary distribution* $\pi = (\pi_i)$ on $\mathcal{S}^w$ satisfying $\pi P = \pi$ [60]. Ergodicity may not hold in the case of a short genomic sequence consisting of systematic repeats of a small number of length-w words.

For a genome $\mathcal{G}$ and a genomic sequence H taken from $\mathcal{G}$, a *genomic signature* for H is a function θ, mapping H to a mathematical structure $\theta(H)$. Ideally, $\theta(H)$ is efficiently computable and can identify sufficiently large substrings that come from $\mathcal{G}$ and accurately identify the origin genome $\mathcal{G}$ of H from a set of genomes by using $\theta(H)$.

Previously, we [36, 37] defined several signatures computed from the structure of the DBC and evaluated these and other signatures, such as the word frequency vector (θ^{wfv}) and the dinucleotides odds ratio signature (θ^{dor}). We studied the behavior of the θ^{dbc} signature and presented associated empirical results [37].

Let $H \in \Sigma^*_{\text{DNA}}$ have length $|H| = n$. Fixing word length $w \geq 1$, we obtain $\mathcal{DB}^w(H)$, with associated $\text{vc}(x_i, H)$ and $\text{ec}((x_i, x_j), H)$. Let $\psi \geq 0$ be an integer *threshold*. Let $E^{\leq\psi} = \{(i, j) \in E \mid \text{ec}((i, j), H) \leq \psi\}$, be the set of edges with counts at most ψ. Then *edge deletion* is the process of deleting edges in $E^{\leq\psi}$ from $\mathcal{DB}^w$, while varying ψ from 0 to $\Xi = \max\{\text{ec}((i, j), H) \mid (i, j) \in E\}$ and deleting edges with tied counts in arbitrary order. As ψ increases from 0 to Ξ, the number of isolated vertices increases from 0 to l. Define the *ordered vertex isolation frequency vector* θ^{ovif} as the l-vector whose i^{th} component is the frequency of the last edge whose deletion isolates vertex labeled with the i^{th} string in lexicographic order. The *de Bruijn chain signature* θ^{dbc} is the $2l$-vector $\hat{\pi}_w \cdot \theta^{ovif}_w / 4^{w-1}$, where $\hat{\pi}_w$ is the estimated stationary distribution for the order-w de Bruijn chain and '$\cdot$' represents vector concatenation. Figure 2 illustrates the construction of the θ^{dbc} signature.

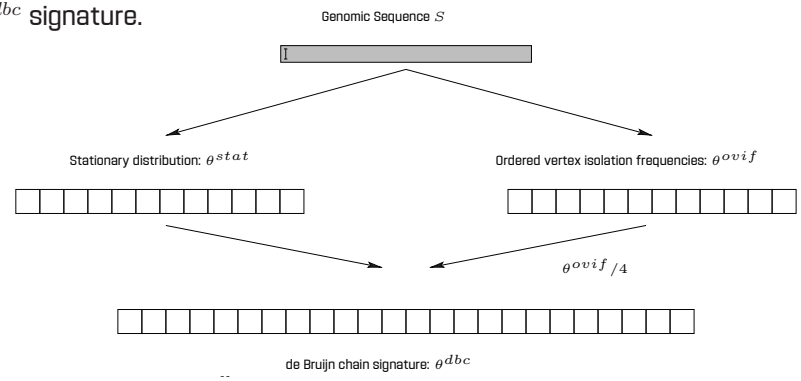


Fig 2. Construction of the θ^{dbc} signature.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

For example, consider the *E. coli K12* genome. The order-2 transition matrix for this sequence is as follows:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC
0.322	0.243	0.187	0.245	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.228	0.291	0.285	0.195	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.237	0.340	0.213	0.210	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.205	0.279
0.237	0.205	0.321	0.237	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.317	0.176	0.321	0.187	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.206	0.332	0.251	0.211	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.114	0.179
0.313	0.204	0.159	0.324	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.249	0.241	0.299	0.209	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.207	0.342	0.176	0.275	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.205	0.213
0.325	0.248	0.127	0.299	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.313	0.209	0.267	0.209	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.259	0.295	0.265	0.181	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.203	0.247

For the given transition matrix, the order-2 stationary distribution is $\langle 0.073 \quad 0.055 \quad 0.051 \quad 0.066 \quad 0.069 \; 0.058 \; 0.074 \; 0.051 \; 0.057 \; 0.082 \; 0.058 \; 0.055 \; 0.045 \; 0.057 \; 0.069 \; 0.073 \rangle$. The corresponding θ_2^{ovif} signature is $\langle 0.325 \; 0.291 \; 0.340 \; 0.324 \; 0.321 \; 0.321 \; 0.332 \; 0.438 \; 0.324 \; 0.342 \; 0.342 \; 0.322 \; 0.325 \; 0.313 \; 0.438 \; 0.323 \rangle$. Therefore, the θ_2^{dbc} signature for this species is the concatenation of the above two vectors.

Our results have indicated that the performance of θ^{dbc} is better than the individual performances of $\hat{\pi}$ and θ^{ovif}. Visualize the components of the θ_w^{ovif} signature as weights on the edges of an edge cover of $\mathcal{DB}^w(H)$. In the edge cover, each vertex remains connected through the strongest edge (edge with highest frequency) incident on it. Figure 3 illustrates this point. For two vector-based signatures θ_1 and θ_2, $d(\theta_1, \theta_2)$

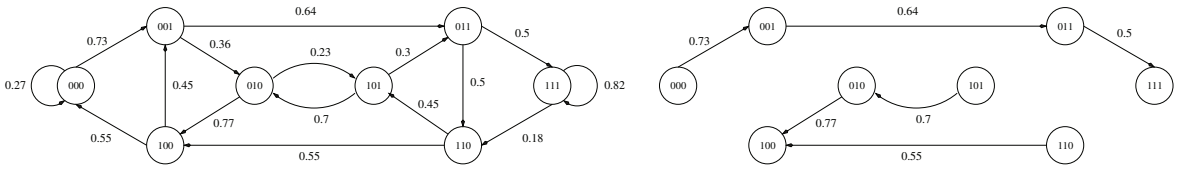


Fig 3. Edge cover example. (a) The binary de Bruijn graph of order 3. (b) The edge cover from which values for individual components of θ_3^{ovif} are taken.

is the L_1 metric in l-dimensional real space and $\mathrm{R}(\theta_1, \theta_2)$ is the Pearson correlation coefficient.

Here, we describe the algorithm used to detect the origin of unknown genomic sequences using the θ^{dbc} signature and study its performance with varying sequence length. We also propose a mathematical framework for characterizing the θ^{dbc} signature (see Theory Section) and explore more properties of this signature while comparing it with the θ^{dor} and θ^{wcv} signatures (see Results Section).

Results

To evaluate the θ^{dbc} signature and to compare its accuracy in sequence origin prediction with that of existing signatures, we used bacterial and eukaryotic genomic sequences. First, we compiled a list L_1 of diverse genomic sequences of various lengths including α-proteobacteria (APB), infectious bacteria, and eukaryotes (Table 1). Second, we collected

http://dx.doi.org/10.21622/ACE.2021.01.1.007

Table 1. List L_1 of genomic sequences in the set of diverse species

Species	Acronym	Sequence length	NCBI identifier
R. leguminosarum	RL	5.1 Mb	NC_008380
E. litoralis	EL	3.1 Mb	NC_007722
M. leprae	ML	3.3 Mb	NC_002677.1
N. meningitidis	NM	2.2 Mb	NC_008767.1
P. falciparum	PF	chr 12, 2.3 Mb	NC_004316.2
P. aeruginosa	PA	6.4 Mb	NC_002516.2
S. pneumoniae	SP	2.1 Mb	NC_008533.1
E.coli	EC	4.7 Mb	NC_000913
C. elegans	CE	chr 1, 15.3 Mb	NC_003279
H. sapiens	HS	chr 1, 228.7 Mb	AC_000044
A. thaliana	AT	chr 4, 18.8 Mb	NC_003075
S. cerevisiae	SC	chr 4, 1.6 Mb	NC_001136

a set of 52 APB genomes including multiple strains of several species to build a collection of genomic sequences derived from closely related species. Of these 52 sequences, the 20 that were used to randomly sample shorter sequences for origin prediction are listed in Table 2 (List L_2). Two databases of θ^{dbc} signatures were constructed; the first database $\mathcal{D}_1^{dbc}$ consisted of the signatures for the sequences in L_1, while the second database $\mathcal{D}_2^{dbc}$ consisted of the signatures for the sequences of the 52 APB, *E. coli*, and the 4

Table 2. List L_2 of genomic sequences in the set of closely-related APB

Species	Sequence length	NCBI identifier
Wolbachia BM	1.1 Mb	NC_006833
R. typhi	1.1 Mb	NC_006142
A. marginale	1.2 Mb	NC_004842
C. pelagibacter	1.3 Mb	NC_007205
A. phagocytophilum	1.5 Mb	NC_007797
B. suis	chr 1, 2.1 Mb	NC_004310
G. bethesdensis	2.7 Mb	NC_008343
P. denitrificans	chr 1, 2.9 Mb	NC_008686
E. litoralis	3.1 Mb	NC_007722
S. alaskensis	3.4 Mb	NC_008048
H. neptunium	3.8 Mb	NC_008358
C. crescentus	4.1 Mb	NC_002696
S. pomeroyi	4.2 Mb	NC_003911
Jannaschia ssp. CCS1	4.4 Mb	NC_007802
R. rubrum	4.4 Mb	NC_007643
N. hamburgensis	4.5 Mb	NC_007964
M. magneticum	5.0 Mb	NC_007626
R. leguminosarum	5.1 Mb	NC_008380
R. palustris	5.6 Mb	NC_008435
M. loti	7.1 Mb	NC_002678

higher eukaryotes from L_1. Similar databases $\mathcal{D}_1^{dor}$ and $\mathcal{D}_2^{dor}$ corresponding to the θ^{dor} signature, and $\mathcal{D}_1^{wcv}$ and $\mathcal{D}_1^{wcv}$ corresponding to the θ^{wcv} signature were also constructed.

Characterization of the accuracy of the θ^{dbc} signature in origin prediction

We tested the ability of the θ^{dbc} signature to distinguish between distant species in [37] using $\mathcal{D}_1^{dbc}$, orders $2-5$, and sample sequence lengths 10 kb, 25 kb, 50 kb, and 100 kb. For each $\langle$order, length$\rangle$ combination, 100 sequences were randomly sampled from each organism in L_1. For each sample X, the vector $\theta_w^{dbc}(X)$ was correlated, using the Pearson correlation coefficient, with all the θ_w^{dbc} vectors in $\mathcal{D}_1$. Accuracy was computed as follows.

For a sample X, the matches to $\theta_w^{dbc}(X)$ were ranked $1, 2, 3, \ldots$ in decreasing order of their correlation coefficients or increasing order of their distances. In a *first hit scenario*, the origin is ranked 1. Depending on the scenario under consideration, the number of first hits per 100 samples is the *accuracy*. For fixed order, we found that the accuracy of origin prediction increases with increasing sample size, reaching 100% first hits at length 100 kb for all species at order 4 (Figure 3 in [37]). Intuitively, a larger sequence encodes more information about the underlying DBC, which leads to the calculation of a θ^{dbc} signature more representative of the origin. The θ^{dbc} signature appeared to be more conserved at order 4 than at other orders. This coincides with the hypothesis behind the application TETRA [22], which also attempts to discover the origin of unknown DNA sequences but does not work well with short sequences. Note that sufficient information about the underlying DBC of order 4 can only be acquired from sequences of size 50 kb or higher under our model; this is not helpful in identifying origins of short DNA sequences. A short sequence contains maximum information about the underlying DBC of order 1. Although the θ_1^{dbc} signature is computable in less time than higher order signatures, it encodes information about mononucleotides only, which is insufficient to accurately predict origin. So, for identification of the origin of short DNA sequences, we use the more accurate and origin-representative order-2 signature θ_2^{dbc}.

Figure 4 shows the distributions of the Pearson correlation coefficients between the θ_2^{dbc} signature of a sample sequence and the θ_2^{dbc} signatures of other sequences in the database including those of the origin of the sample sequence. For each species on the x-axis, there are 2 box and whisker plots generated as follows. 100 samples of length 50 kb each are randomly sampled from the genome of each species. The correlation of the θ_2^{dbc} signature of each sample with the θ_2^{dbc} signature of its origin is binned separately from its correlations with the θ_2^{dbc} signatures of all other organisms. The distribution of numbers in each bin is represented by a box and whisker plot along the y-axis. The smaller box plots with medians close to 1 and small ranges between the first quartile and the third quartile represent the distribution of correlations of signatures of sample sequences with the signatures of their origin. The larger box plots with large ranges between their first and third quartiles and smaller medians represent the distribution of correlations with species other than the origin. These data demonstrate that the θ^{dbc} signature retains features unique to each organism and can differentiate between the origin and other species. It is highly conserved within a genome and differs between genomes.

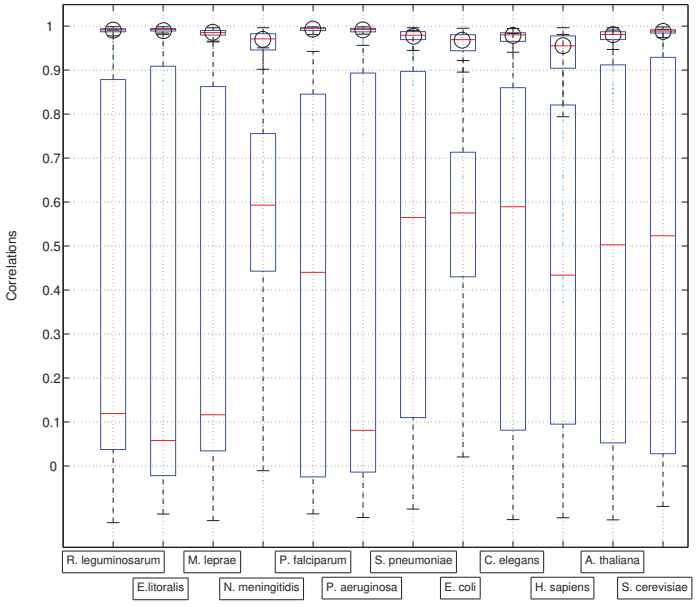


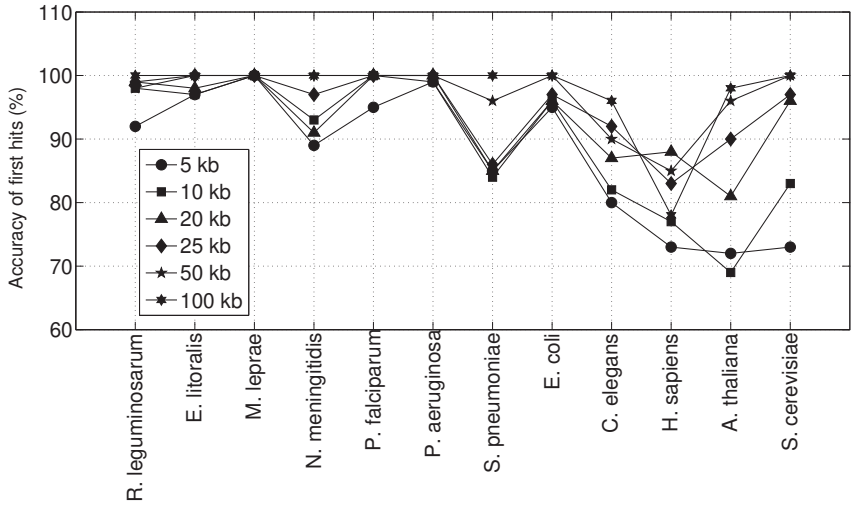
Fig 4. Performance of θ_2^{dbc}. The 12 species are on the x-axis. The small box and whisker plots near the top (with associated circles) represent the distribution of correlations of θ_2^{dbc} signatures of the 100 samples with the θ_2^{dbc} signatures of their respective origins. The larger box and whisker plots represent the distribution of correlations with θ_2^{dbc} signatures of other genomes.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

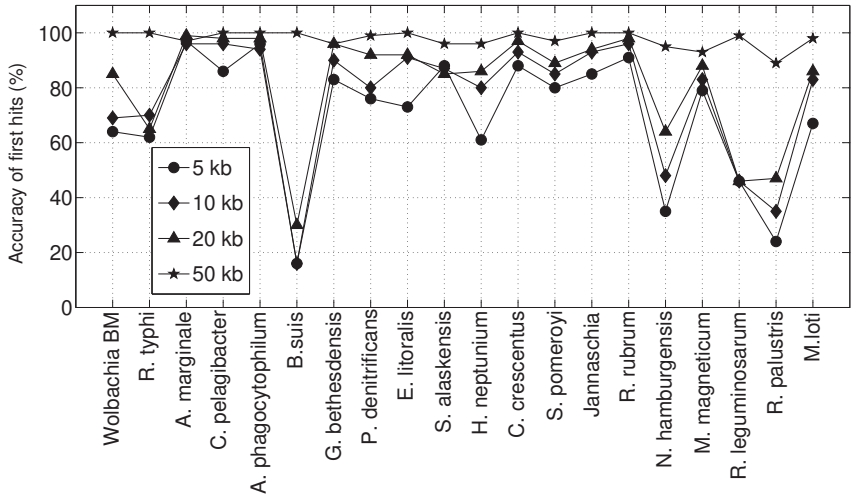
Second, we tested the ability of θ^{dbc} to distinguish between genomic sequences from closely-related species and different strains of the same species using $\mathcal{D}_2$ (57 signatures) as the database, and the 20 sequences in L_2 for sampling. Results are presented for sample sequences of lengths 10 kb and 50 kb in Figure 5 in [37]. We observed that the order-2 θ^{dbc} signature is better at distinguishing between closely related species than θ^{dbc} signatures of higher order.

The accuracy of θ_2^{dbc} for both test cases is summarized in Figure 5. For list L_1, a median accuracy greater than 90% is achieved for sequences as short as 5 kb. Median accuracy increases steadily with sample size and is 100% at a sample length of 50 kb.

(a)



(b)



(c)

Sample sequence length	Median accuracy of first hits	
	List in Table 1	List in Table 2
5 kb	90.5	77.5
10 kb	94.5	84
20 kb	96	93
25 kb	97	-
50 kb	100	99
100 kb	100	-

Fig 5. Summary of accuracy of first hits of θ_2^{dbc} in both experiments using (a) Species in L_1, (b) Species in L_2. (c) Listing of median first hit accuracies of origin prediction for various sample sequence lengths using θ_2^{dbc}. The hyphens indicate placeholders for entries that were computed not for 100 samples, but for a lesser number of samples, and hence, are not shown here.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

For the human genome, the θ_2^{dbc} signature consistently does not perform well. This issue is addressed in the Comparison of Performances Section, where we compare different signatures and discuss conservation of specific features in each genome. Distinguishing between closely-related species is a harder task than distinguishing between diverse-species. The signature must capture subtle differences at a much finer scale between two closely-related sequences to be able to tell them apart. Hence, the reduced accuracy in case of list L_2 is expected. In case of list L_1, a median accuracy greater than 84% is achieved for sequences of length 10 kb, and improves to almost 100% on increasing the sample sequence size to 50 kb. We note that sample sequences of length 20 kb are sufficient to predict the origin with reasonably high accuracy.

Comparison of performances of θ^{dbc}, θ^{dor}, and θ^{wcv} signatures

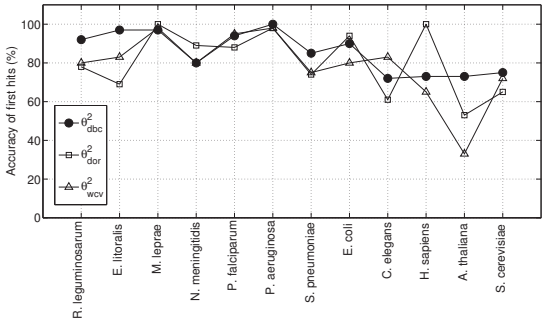
We compared the accuracy of the three signatures θ^{dbc}, θ^{dor}, and θ^{wcv} in predicting the origin of short DNA segments. The same methods and terminologies as described in Section have been used. Order 2 signatures were used for several reasons. In Section , we found order-2 DBCs to be most representative of the origin in the case of short sequences and the corresponding θ_2^{dbc} to be more quickly computable than higher order signatures. Also, the θ^{dor} signature has an underlying order of 2, hence, using the same order for its competitors is fair.

First, the ability of all three signatures to distinguish between highly separated species was tested using the list L_1 for sampling and $\mathcal{D}_1$ databases for each type of signature. Shorter sequence samples of lengths 5 kb, 10 kb, and 20 kb were used. Figure 6(a), (b), and (c) illustrate the results. 100 subsequences were randomly sampled from each of the 12 diverse species on the x-axis. All three signatures were computed using each sample and correlated to their respective $\mathcal{D}_1$ databases of signatures. The accuracy of first hits are recorded on the y-axis.

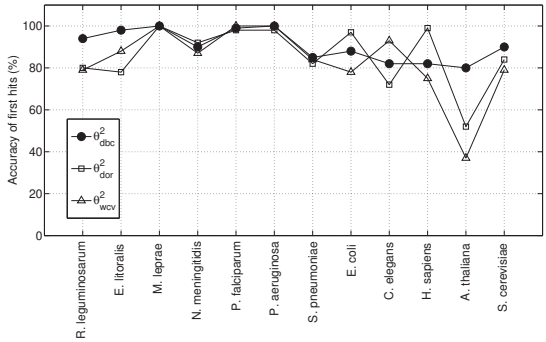
Observe that the θ_2^{dbc} signature outperforms the θ^{dor} signature for all sequence lengths by demonstrating better accuracy in the case of $8/12$, $9/12$, and $8/12$ species for sequence lengths 5 kb, 10 kb, and 20 kb, respectively. The θ_2^{dbc} signature also outperforms the θ_2^{wcv} signature for all sequence lengths by demonstrating better accuracy in the case of $9/12$, $10/12$, and $11/12$ species for sequence lengths 5 kb, 10 kb, and 20 kb, respectively. The only genomes for which θ_2^{dbc} consistently demonstrates worse accuracy than θ^{dor} are NM, EC, and HS. Of particular interest is the HS (*Homo sapiens*) genome, where the θ^{dor} signature appears to be very well conserved demonstrating almost 100% accuracy irrespective of the sample sequence length. In the rest of the genomes (RL, EL, ML, PF, PA, SP, CE, AT, SC), the θ_2^{dbc} signature is better conserved than the θ^{dor} signature. Compared with the θ_2^{wcv} signature, the θ_2^{dbc} signature consistently performs worse only in case of CE. For all other species, the accuracy of the θ_2^{dbc} signature is better than or equal to that of the θ_2^{wcv} signature. Consider Figure 7. In Figure 7(a), for each species on the x-axis, the y-axis plots the number of samples out of 100 for each sequence length, where θ_2^{dbc} outperformed the θ^{dor} signature in predicting the origin of the sample. Observe that with decreasing sequence length, the relative predictive accuracy of θ_2^{dbc} increases and is an advantage over that of the θ^{dor} signature. The exceptions are the three species pointed out above where θ^{dor} is more well-conserved than θ_2^{dbc}. The same behavior is repeated in the case of the comparison between prediction accuracies of θ_2^{dbc} and θ_2^{wcv} in Figure 7(b) with CE being the only exception.

Next, we compared the abilities of the three signatures to distinguish between closely-related species while using the list L_2 for sampling and the $\mathcal{D}_2$ databases for each type of signature. Short sequence samples of lengths 5 kb, 10 kb, and 20 kb were used. Figure 6(d), (e), and (f) illustrate the results. The same method was followed as in the previous case of diverse species. The accuracy of first hits are recorded on the y-axis.

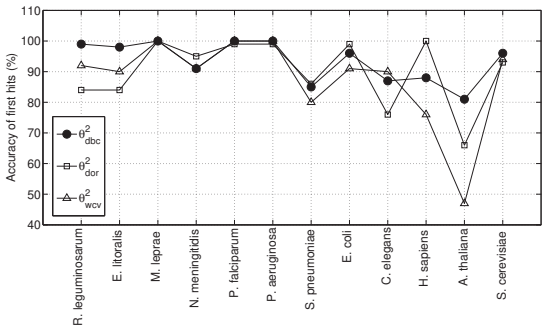
http://dx.doi.org/10.21622/ACE.2021.01.1.007



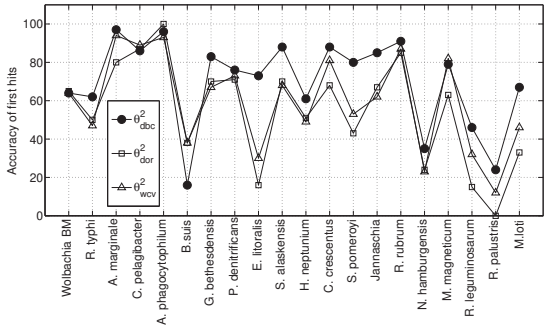
(a) 5 kb sample sequences from L_1 matched against $\mathcal{D}_1^*$



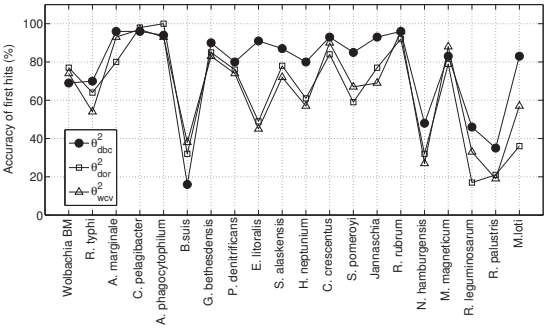
(b) 10 kb sample sequences from L_1 matched against $\mathcal{D}_1^*$



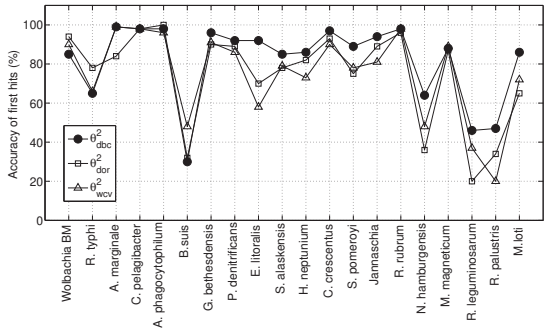
(c) 20 kb sample sequences from L_1 matched against $\mathcal{D}_1^*$



(d) 5 kb sample sequences from L_2 matched against $\mathcal{D}_2^*$



(e) 10 kb sample sequences from L_2 matched against $\mathcal{D}_2^*$

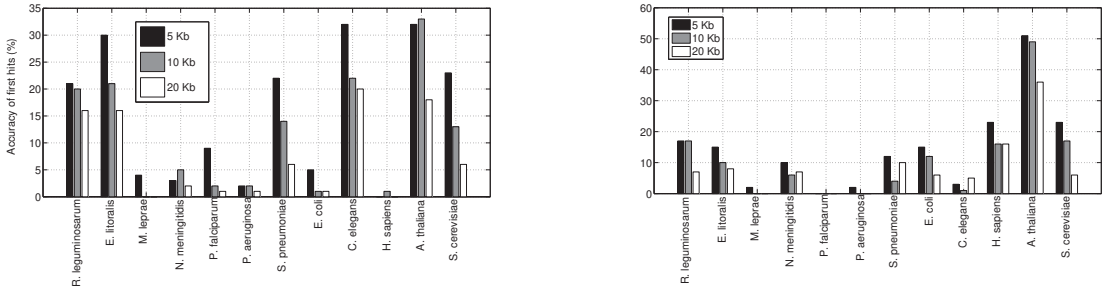


(f) 20 kb sample sequences from L_2 matched against $\mathcal{D}_2^*$

Fig 6. Accuracy of first hits of θ_2^{dbc}, θ^{dor}, and θ_2^{wcv} signatures. 100 Sample sequences of lengths (a) 5 kb, (b) 10 kb, and (c) 20 kb have been used from each species from the list L_1 of diverse species on the x-axis. 100 Sample sequences of lengths (d) 5 kb, (e) 10 kb, and (f) 20 kb have been used from each species from the list L_2 of closely related APB on the x-axis. The y-axis represents the number of first hits out of 100. The legends in the plots indicate specific data for each signature.

The database, in this case, contains 52 species from the same family (APB) and 5 other diverse species. Figure 6(d), (e), and (f) illustrate that the θ_2^{dbc} signature outperforms both θ^{dor} and θ_2^{wcv} signatures in the case of all sequence lengths with better predictive accuracy for $15/20$ species against the θ^{dor} signature and an average better accuracy of $16.33/20$ species against the θ_2^{wcv} signature. The θ^{dor} signature appears consistently more well-conserved than the θ_2^{dbc} signature in the case of *Wolbachia*. In the comparison between the θ_2^{dbc} and θ_2^{wcv} signatures, the θ_2^{dbc} signature is consistently at least as well conserved as its competitor in all species but that of *B. suis*. Even in the case of closely-related species, the relative accuracy of the θ_2^{dbc} signature increases with decreasing

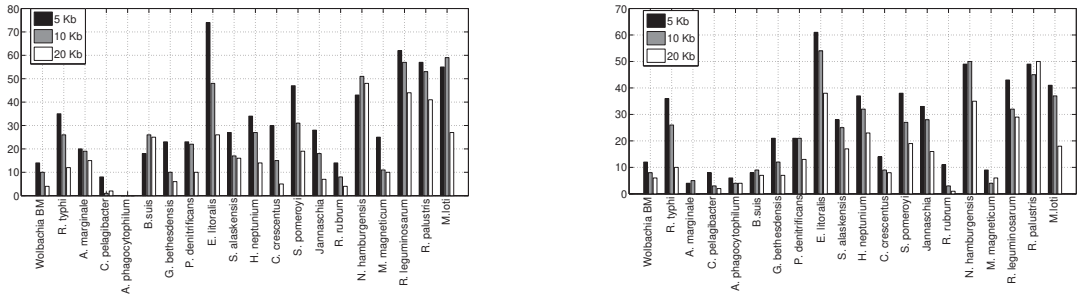
http://dx.doi.org/10.21622/ACE.2021.01.1.007



(a) θ_2^{dbc} vs. θ^{dor} (b) θ_2^{dbc} vs. θ_2^{wcv}

Fig 7. Comparison of relative accuracies of (a) θ_2^{dbc} and θ_2^{dor} and (b) θ_2^{dbc} and θ_2^{wcv} for sequence lengths 5 kb, 10 kb, and 20 kb. For each species on the x-axis, the y-axis represents the number of samples out of 100 where the θ_2^{dbc} signature outperforms its competitor.

sequence length as is demonstrated by the data in Figure 8.



(a) θ_2^{dbc} vs. θ^{dor} (b) θ_2^{dbc} vs. θ_2^{wcv}

Fig 8. Comparison of relative accuracies of (a) θ_2^{dbc} and θ^{dor} and (b) θ_2^{dbc} and θ_2^{wcv} for sequence lengths 5 kb, 10 kb, and 20 kb randomly sampled from APB. For each species on the x-axis, the y-axis represents the number of samples out of 100 where the θ_2^{dbc} signature outperforms its competitor.

A more formal statistical comparison of the accuracy of these signatures using the Wilcoxon signed rank test is summarized in Table 3. While differentiating between diverse species, for sequence length $<$ 20 kb, the median accuracy of θ_2^{dbc} is greater than the median accuracy of θ_2^{wcv} to a statistically significant extent. While differentiating between closely-related species, the same can be said for sequence length $<$ 5 kb. θ_2^{dbc}

and θ^{dor} are more evenly matched than θ_2^{dbc} and θ_2^{wcv} in case of diverse species, whereas in case of closely-related species, the θ_2^{dbc} signature clearly outperforms the θ^{dor} signature to a statistically significant extent.

For the order-2 signatures above, Figure 9 summarizes the median accuracy of prediction of first hits in the case of both lists L_1 and L_2 and varying sequence lengths of 5 kb, 10 kb, and 20 kb. Observe that in all cases, the θ_2^{dbc} signature outperforms the θ^{dor} signature, which in turn outperforms the θ_2^{wcv} signature.

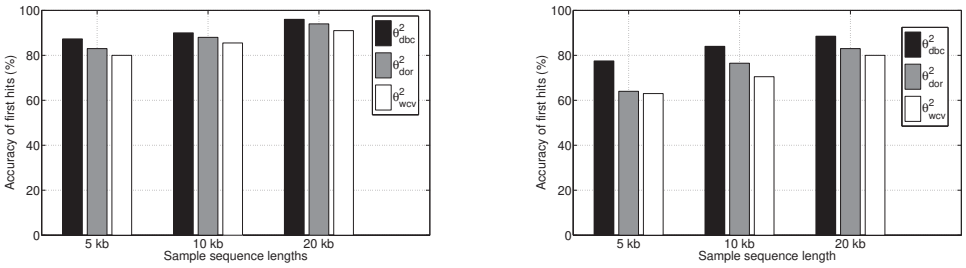
Combining the powers of θ_2^{dbc} and θ^{dor}

In the Comparison of Performances Section, we demonstrated that, in predicting the origin of an unknown DNA sequence, the θ_2^{dbc} signature has greater accuracy than the

http://dx.doi.org/10.21622/ACE.2021.01.1.007

Table 3. Wilcoxon signed rank test results comparing the accuracies of different signatures. For instance, for list L_1 and signatures θ_2^{dbc} and θ^{dor}, X is an accuracy vector of length 12 for θ_2^{dbc} and Y is an accuracy vector of length 12 for θ^{dor}. The null hypothesis being tested here is H_0 : median_accuracy(Sign 1)=median_accuracy(Sign 2). The alternate hypothesis is H_1 : median_accuracy(Sign 1)>median_accuracy(Sign 2).

List	Sequence length	Sign 1 (X)	Sign 2 (Y)	Mean difference	Signed rank	p-value (One sided)	Accept H_1? ($\alpha = 0.05$)
L_1	5 kb	θ_2^{dbc}	θ^{dor}	+3.08	37	0.0314	Yes
L_1	5 kb	θ_2^{dbc}	θ_2^{wcv}	+4.92	58	0.0052	Yes
L_1	10 kb	θ_2^{dbc}	θ^{dor}	+2.75	34	0.0681	No
L_1	10 kb	θ_2^{dbc}	θ_2^{wcv}	+2.66	31	0.0371	Yes
L_1	20 kb	θ_2^{dbc}	θ^{dor}	+2.08	25	0.1379	No
L_1	20 kb	θ_2^{dbc}	θ_2^{wcv}	+2.67	32	0.011	Yes
L_2	5 kb	θ_2^{dbc}	θ^{dor}	+8.4	168	0.0009	Yes
L_2	5 kb	θ_2^{dbc}	θ_2^{wcv}	+7.3	146	0.0017	Yes
L_2	10 kb	θ_2^{dbc}	θ^{dor}	+7.6	152	0.0023	Yes
L_2	10 kb	θ_2^{dbc}	θ_2^{wcv}	+1.15	23	0.3372	No



(a) List in Table 1 (b) List in Table 2

Fig 9. Comparison of median first hit accuracies of origin prediction for θ_2^{dbc}, θ^{dor}, and θ_2^{wcv} signatures. The x-axis represents sample sequence lengths. The y-axis represents accuracy of first hits.

θ^{dor} and θ_2^{wcv} signatures. The objective of this work is not to introduce yet another genomic signature. We are interested in exploring aspects of construction of a genome that make it unique. So far, we have been successful in discovering some such aspects through the θ^{dbc} signature.

To reiterate our observations from Section , the θ_2^{dbc} demonstrates high accuracy when sample sequences of length 20 kb or higher are available, both in differentiating between far-away species and closely-related species. It is the case when much shorter samples are present and the accuracy drops. To test whether an even greater accuracy of origin prediction for short sequences can be achieved, we combined the θ_2^{dbc} and θ^{dor} signatures. We tried three different methods of doing the above. We concatenated the two signatures into one vector and used Pearson correlations to determine the closest species. This method works no better than using individual θ_2^{dbc} signatures. Working with the sum of the Pearson correlation distance and the normalized L_1-distance, separately computed, did not yield better results either. However, using the product of the Pearson correlation distance and the normalized δ-distance, separately computed, appears to produce a better accuracy than using the θ_2^{dbc} signature alone, in the case of differentiating between far-away species as observed in Figure 10.

However, the same method does not demonstrate substantially higher accuracy than the θ_2^{dbc} signature in differentiating between closely-related species. We make

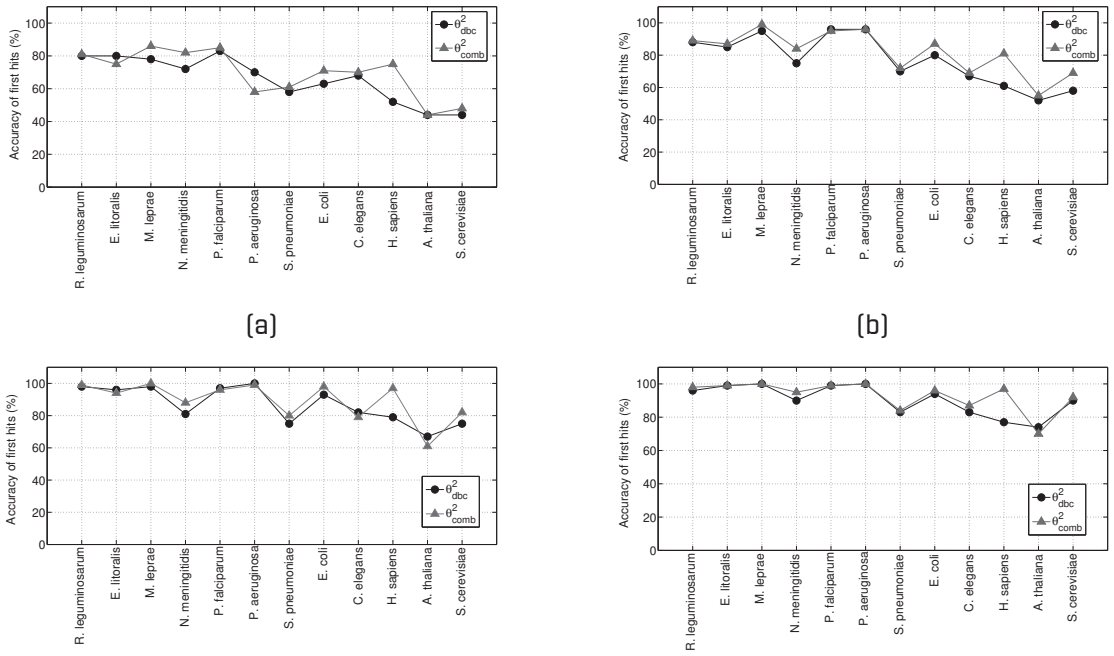


Fig 10. Comparison of the accuracies of the θ_2^{dbc} signature and the combination signature θ_2^{comb} of θ_2^{dbc} and θ^{dor} in predicting origins of unknown short sequences from L_1. Sequences of lengths (a) 1 kb, (b) 2 kb, (c) 5 kb, and (d) 10 kb have been used.

this observation based on the results in in Figure 11. In fact, in this case, accuracy drops to less than 25% for most species when the sample sequence length is approximately 1 kb, which is why results corresponding to such short sequences are not shown.

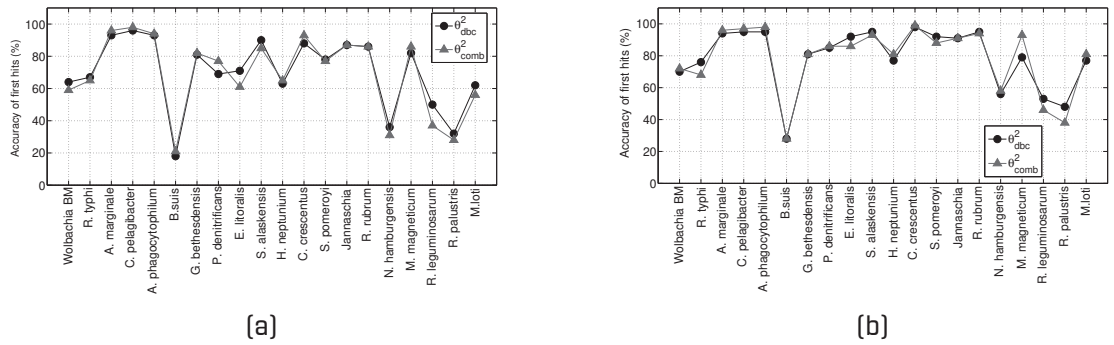


Fig 11. Comparison of the accuracies of the θ_2^{dbc} signature and the combination signature θ_2^{comb} of θ_2^{dbc} and θ^{dor} in predicting origins of unknown short sequences from L_2. Sequences of lengths (a) 5 kb and (b) 10 kb have been used.

Relationship between genome size and accuracy of origin prediction

Next, we explored pairwise relationships between genome size and efficiency of first hits of the θ_2^{dbc} signature using sequence samples of length 10 kb. Figure 12 presents relevant scatter plots. Plot (a) is for 11 out of 12 genomes in L_1 (the human genome was not used as it was an outlier that disrupted the otherwise observed correlations, because of its

http://dx.doi.org/10.21622/ACE.2021.01.1.007

large size), while plot (b) is for the 20 APB. Observe that the efficiency of θ^{dbc} is negatively correlated with genome size in both sets, using Pearson correlation coefficients.

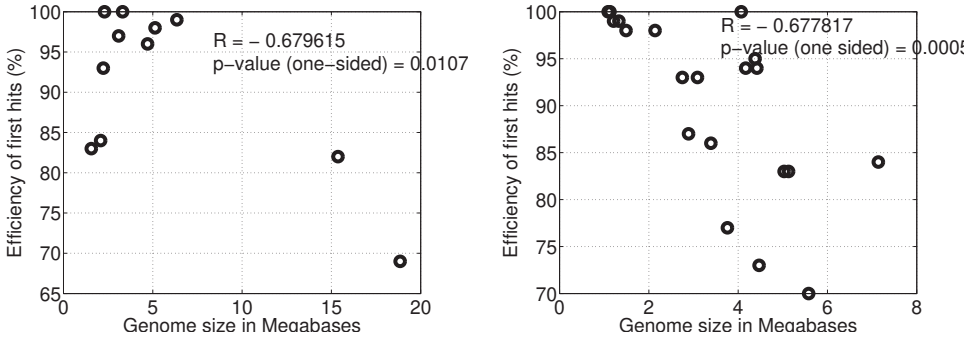


Fig 12. Relationships between genome size and efficiency of first hits of θ_2^{dbc}. (a) and (b) demonstrate results for the 11 species in Table 1 and the 20 APB, respectively.

Theory

In this section, we build a theoretical framework to analyze distances between θ_2^{dbc} signatures in terms of the parameters of the DBCs generating them. The results for higher order θ^{dbc} signatures can be derived in a similar manner as θ_2^{dbc} signatures. Let $\mathcal{DC}$ be an ergodic, order-2 DBC. Let H be a sequence generated by $\mathcal{DC}$, where $|H| = n$. If $x_i, x_j \in \mathcal{S}^2$, the probability of transition from state x_i to state x_j is given by $p_{i,j}$, while the stationary probability for x_i is $\hat{\pi}_{x_i}$.

As defined in the Preliminaries Section, the θ_2^{dbc} signature is a concatenation of the $\hat{\pi}_2$ signature and the $\theta_2^{ovif}/4$ signature. First we develop a framework for characterizing $\hat{\pi}_2$.

Framework for bounding the distance between $\hat{\pi}_2$ signatures derived from sequences generated by the same DBC

Let H be a long DNA sequence generated by an order-w DBC with irreducible transition matrix P and stationary distribution $\hat{\pi}_w(H)$. Let h be a much shorter subsequence of H with transition matrix P' and stationary distribution $\hat{\pi}_w(h)$. Assuming that P' is irreducible, P' is a perturbed form of P. When P and P' are close, the distance between $\hat{\pi}_w(H)$ and $\hat{\pi}_w(h)$ is very small and can be bound.

Solan and Vieille [61] have defined a measure of closeness of P' to P. They define ζ as $\zeta_P = \min_{C \subset \mathcal{S}^w} \sum_{s \in C} \hat{\pi}_w(H) P(\bar{C}|s)$. They state that P' is $((\epsilon, b))$-close to P if for all pairs of states $s, t \in \mathcal{S}^w$, $|1 - P'(t|s)/P(t|s)| \leq b$ whenever (a) $\hat{\pi}_s^w(H) P(t|s) \geq \epsilon\zeta_P$ or (b) $\hat{\pi}_s^w(H) P'(t|s) \geq \epsilon\zeta_P$. Let $L = \sum_{i=1}^{|\mathcal{S}^w|-1} \binom{|\mathcal{S}^w|}{i} i^{|\mathcal{S}^w|}$. Then, if $b \in (0, 1/2^{|\mathcal{S}^w|})$ and $\epsilon \in \left(0, \frac{b(1-b)}{L|\mathcal{S}^w|^4}\right)$, for every transition matrix P' that is (ϵ, b)-close to P, (i) P' is irreducible and (ii) Its stationary distribution $\hat{\pi}_w(h)$ satisfies $|1 - \hat{\pi}_s^w(h)/\hat{\pi}_s^w(H)| \leq 18bL$. A detailed proof of the above lemma can be found in Solan and Vieille [61].

http://dx.doi.org/10.21622/ACE.2021.01.1.007

From the above discussion, it is clear that for a genomic sequence H generated by order-w DBC $\mathcal{DC}$ and its much smaller subsequence h, the stationary distribution of $\mathcal{DC}$ can be accurately represented by $\hat{\pi}_w(H)$ and closely approximated by $\hat{\pi}_w(h)$. Therefore, the estimated stationary distribution of the DBC that generates a genomic sequence, can serve as a genomic signature.

Our results [37] (not shown here) suggest that $\theta_w^{wfv}(H) \approx \hat{\pi}_w(H)$, and $\hat{\pi}_w(h) \approx \hat{\pi}_w(H)$, while $\theta_w^{wfv}(h)$ might not display such similarity to either $\theta_w^{wfv}(H)$ or $\hat{\pi}_w(H)$. This property is conserved for a wide range of lengths of h (tested for ≥ 5 kb). In Theorem 1, we bound the distance between the stationary distributions derived from the transition matrices of sequences generated by the same DBC. First, we prove the following lemma.

Lemma 1. *Let H be a genomic sequence of length n generated by an order 2 DBC with underlying stationary distribution π. Assume that the number of occurrences of a dinucleotide x has a Poisson distribution with mean $n\pi_x$. Let $\hat{\pi}_x$ be the random variable representing the stationary probability of x. Fix $\tau > 0$ and $T = n\tau$. Then*

$$\Pr\left[|\hat{\pi}_x - \mathsf{E}\left[\hat{\pi}_x\right]| > \tau\right] < \mathcal{L}^{\pi}(x) + \mathcal{U}^{\pi}(x), \text{ where}$$

$$\mathcal{L}^{\pi}(x) = \exp\left(\frac{-T^2}{2n\pi_x}\right) \text{ and } \mathcal{U}^{\pi}(x) = \left(\frac{\mathsf{e}^{\frac{T}{n\pi_x}}}{\left(1+\frac{T}{n\pi_x}\right)^{1+\frac{T}{n\pi_x}}}.\right)^{n\pi_x}.$$

Proof. Let X_x be the random variable representing the number of occurrences of the dinucleotide x. Then X_x can be expressed as a sum of $n-1$ indicator random variables, each representing the occurrence of x at a given position in the sequence. In particular, $X_x = \sum_{i=1}^{n-1} X_x(i)$, where $\Pr\left[X_x(i) = 1\right]$ is equal to π_x for all i, and $\mathsf{E}\left[X_x\right] \approx n\pi_x$. Now,

$$\Pr\left[|\hat{\pi}_x - \mathsf{E}\left[\hat{\pi}_x\right]| > \tau\right] = \Pr\left[|X_x - \mathsf{E}\left[X_x\right]| > n\tau\right] = \Pr\left[|X_x - \mathsf{E}\left[X_x\right]| > T\right].$$

Since X_x can be expressed as a sum of independent indicator random variables, Chernoff's bounds [62] are applicable. For the lower tail of the above probability, the applicable Chernoff bound [62] is

$$\Pr\left[X_x < (1-\delta)\mu\right] < \mathsf{e}^{\frac{-\mu\delta^2}{2}}, \text{ where } \mu = \mathsf{E}\left[X_x\right].$$

Using $\Pr\left[X_x - \mathsf{E}\left[X_x\right] < -T\right] = \Pr\left[X_x < n\pi_x - T\right] = \Pr\left[X_x < (1-\delta)n\pi_x\right]$ we get, $n\pi_x - T = (1-\delta)n\pi_x$ or $\delta = \frac{T}{n\pi_x}$. Therefore, the lower tail probability is bounded as follows:

$$\Pr\left[X_x - \mathsf{E}\left[X_x\right] < -T\right] < \exp\left(\left(-n\pi_x/2\right) \cdot \left(T/n\pi_x\right)^2\right) = \exp\left(-T^2/2n\pi_x\right) = \mathcal{L}^{\pi}(x).$$

For the corresponding upper tail of the probability, the applicable Chernoff's bound [62] is

$$\Pr\left[X_x > (1+\delta)\mu\right] < \left(\mathsf{e}^{\delta}/(1+\delta)^{1+\delta}.\right)^{\mu}$$

Using $\Pr\left[X_x - \mathsf{E}\left[X_x\right] > T\right] = \Pr\left[X_x > n\pi_x + T\right] = \Pr\left[X_x > (1+\delta)n\pi_x\right]$ we get $n\pi_x + T = (1+\delta)n\pi_x$ or $\delta = \frac{T}{n\pi_x}$. Therefore, the upper tail probability is bound as follows:

$$\Pr\left[X_x - \mathsf{E}\left[X_x\right] > T\right] < \left(\mathsf{e}^{\frac{T}{n\pi_x}} \Big/ \left(1+T/n\pi_x\right)^{1+T/n\pi_x}\right)^{n\pi_x} = \mathcal{U}^{\pi}(x).$$

Combining the two tail probabilities proves the Lemma.

—

Theorem 1. *Let H_1 and H_2 be genomic sequences of length n independently generated by the same order 2 DBC with underlying stationary distribution π. Let $\hat{\pi}^1$ and $\hat{\pi}^2$ be the order 2 stationary distributions derived from the respective transition matrices of H_1 and H_2. Assume that the number of occurrences of a dinucleotide x has a Poisson distribution with mean $n\hat{\pi}_x$. Then for $\tau > 0$ and $T = n\tau$,*

http://dx.doi.org/10.21622/ACE.2021.01.1.007

$$\Pr\left[d\left(\hat{\pi}^1,\hat{\pi}^2\right)>32\tau\right] \quad < \quad 2\cdot\sum_{x\in\mathcal{S}^2}\left(\mathcal{L}^{\pi}(x)+\mathcal{U}^{\pi}(x)\right).$$

Proof. Using the bound for the stationary distribution of each dinucleotide as derived in Lemma 1 and applying the union bound we have

$$\Pr\left[\left|\hat{\pi}^1-\mathsf{E}\left[\hat{\pi}^1\right]\right|>16T/n\right] \quad \leq \quad \sum_{x\in\mathcal{S}^2}\left(\mathcal{L}^{\pi}(x)+\mathcal{U}^{\pi}(x)\right) \text{ and}$$

$$\Pr\left[\left|\hat{\pi}^2-\mathsf{E}\left[\hat{\pi}^2\right]\right|>16T/n\right] \quad \leq \quad \sum_{x\in\mathcal{S}^2}\left(\mathcal{L}^{\pi}(x)+\mathcal{U}^{\pi}(x)\right).$$

The expected value of $\hat{\pi}_x$ for any x is the same in both sequences H_1 and H_2. Therefore, $d\left((\hat{\pi}^1-\mathsf{E}\left[\hat{\pi}^1\right]),(\hat{\pi}^2-\mathsf{E}\left[\hat{\pi}^2\right])\right)=d\left(\hat{\pi}^1,\hat{\pi}^2\right)$. Using the union bound we get,

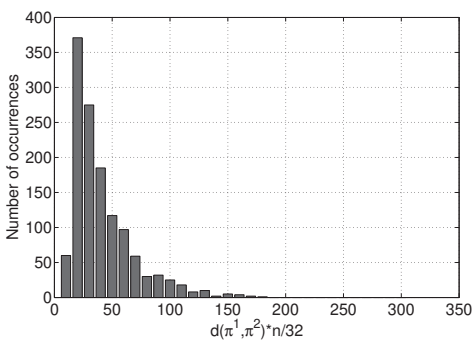
$$\begin{aligned}\Pr\left[d\left(\hat{\pi}^1,\hat{\pi}^2\right)>32\tau\right] &= \Pr\left[d\left(\hat{\pi}^1-\mathsf{E}\left[\hat{\pi}^1\right],\hat{\pi}^2-\mathsf{E}\left[\hat{\pi}^2\right]\right)>32T/n\right]\\ &\leq \Pr\left[|\hat{\pi}^1-\mathsf{E}\left[\hat{\pi}^1\right]|>16T/n\right]\\ &+ \Pr\left[|\hat{\pi}^2-\mathsf{E}\left[\hat{\pi}^2\right]|>16T/n\right]\\ &= 2\cdot\sum_{x\in\mathcal{S}^2}\left(\mathcal{L}^{\pi}(x)+\mathcal{U}^{\pi}(x)\right).\end{aligned}$$

The quantity τ is indicative of the amount of separation that can exist in between two signatures with high probability. The R.H.S. in Theorem 1 is the probability that the separation exceeds a linear function of τ.

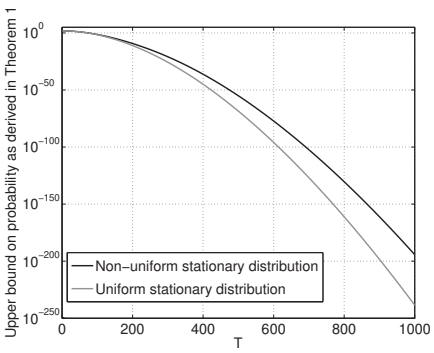
We study the nature of the bound in Theorem 1 as follows. In Figure 13(a), we have plotted the distribution of $T=n\tau$ values using τ values computed from L_1 distances between sequences sampled from the same organism. Sequences of size 10 kilobases were used. A set of genomic sequences were randomly selected. From each genomic sequence 100 pairs of subsequences were independently sampled at random, their stationary distributions were estimated, and the L_1 distance between each pair of stationary distributions was recorded. τ and subsequently, T values were computed from this distance and their distribution plotted as in Figure 13(a). In Figure 13(b), the theoretical bounds are simulated for different values of T and the upper bounds on probability are plotted using $n=10000$ and both uniform and non-uniform stationary distributions. When $T>160$ (approximately), the corresponding probability of separation is very low. This illustrates the strong connection between the bound and real data from genomes.

Framework for bounding the distance between θ_2^{ovif} signatures derived from sequences generated by the same DBC

We begin by characterizing the distribution of the transition probability between two states. Let t be the transition $\sigma_1\ldots\sigma_w\rightarrow\sigma_2\ldots\sigma_{w+1}$. Let X and Y be random variables denoting the number of occurrences of $\alpha=\sigma_1\sigma_2\ldots\sigma_{w+1}$ and $\beta=\sigma_1\ldots\sigma_w$ respectively



(a)



(b)

Fig 13. (a) Plot of distribution of T values computed using τ values taken from L_1 distances between stationary distributions of sequences from the same genome. The L_1 distance between each pair was equated to 32τ. τ, and subsequently T, were derived and the distribution of T values was computed and plotted. Note that approximately $T > 150$ indicates a large and unlikely separation between $\hat{\pi}$ signatures of sequences generated by the same DBC. (b) Plot of upper bounds of separation between stationary distributions of sequences from the same DBC using the theoretical expression derived in Theorem 1.

in a sequence H. The random variable Z representing the estimated probability of the transition t is

$$Z = \begin{cases} X/Y & \text{if } Y \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 2 presents an upper bound on the probability of a specified separation between the frequency of a given transition t and its expected value.

Lemma 2. *Assume, for α and β as described above, that, given an occurrence of β, the occurrence of α is binomially distributed with parameter π_α/π_β. Let a sequence H of length n be given along with a transition t represented by the random variable Z as defined above. Then for $\tau > 0$,*

$$\mathsf{Pr}\left[|Z/4 - \mathsf{E}\left[Z/4\right]| \geq \tau\right] < \mathcal{L}^{ovif}(\beta) + \mathcal{U}^{ovif}(\beta), \text{ where}$$

$$\mathcal{L}^{ovif}(\beta) = \mathsf{e}^{-n\pi_\beta}\left(\exp\left(\exp\left(-8\tau^2\frac{\pi_\beta}{\pi_\alpha}\right)(n\pi_\beta)\right) - 1\right) \text{ and}$$

$$\mathcal{U}^{ovif}(\beta) = \mathsf{e}^{-n\pi_\beta}\left(\exp\left(\left(\frac{\mathsf{e}^{\frac{4\tau\pi_\beta}{\pi_\alpha}}}{\left(1+\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^{1+\frac{4\tau\pi_\beta}{\pi_\alpha}}}\right)^{\frac{\pi_\alpha}{\pi_\beta}}(n\pi_\beta)\right) - 1\right).$$

Proof. Recall that the θ_2^{ovif} signature is scaled by 4 to maintain similar orders of magnitude as the $\hat{\pi}_w$ signature. Let X_α and X_β be random variables representing the number of occurrences of strings α and β respectively in H.

$$\begin{aligned}
\mathsf{Pr}\left[|Z/4 - \mathsf{E}\left[Z/4\right]| \geq \tau\right] &= \mathsf{Pr}\left[\left|\frac{X_\alpha}{4X_\beta} - \mathsf{E}\left[\frac{X_\alpha}{4X_\beta}\right]\right| \geq \tau\right] \\
&= \sum_{c=1}^{\infty}\mathsf{Pr}\left[\left|\frac{X_\alpha}{4c} - \frac{1}{4}\mathsf{E}\left[\frac{X_\alpha}{X_\beta}\right]\right| \geq \tau \mid X_\beta = c\right]\cdot\mathsf{Pr}\left[X_\beta = c\right] \\
&= \sum_{c=1}^{\infty}\mathsf{Pr}\left[\left|X_\alpha - \frac{c\pi_\alpha}{\pi_\beta}\right| \geq 4\tau c \mid X_\beta = c\right]\cdot\frac{\mathsf{e}^{-n\pi_\beta}(n\pi_\beta)^c}{c!}.
\end{aligned}$$

Since X_α can be represented as a sum of independent indicator random variables with $\mathsf{E}\left[X_\alpha\right] = c\pi_\alpha/\pi_\beta$, Chernoff bounds [62] are applicable to the probability $\mathsf{Pr}\left[\left|X_\alpha - \frac{c\pi_\alpha}{\pi_\beta}\right| \geq 4\tau c \mid X_\beta = c\right]$. Consider the lower tail probability

$$\mathsf{Pr}\left[X_\alpha - \frac{c\pi_\alpha}{\pi_\beta} \leq -4\tau c \mid X_\beta = c\right] = \mathsf{Pr}\left[X_\alpha \leq \frac{c\pi_\alpha}{\pi_\beta} - 4\tau c \mid X_\beta = c\right].$$

Using Chernoff's lower tail bounds as described in the proof of Lemma 1 with $\mu = \mathsf{E}\left[X_\alpha\right]$ and using $\frac{c\pi_\alpha}{\pi_\beta} - 4\tau c = (1-\delta)\frac{c\pi_\alpha}{\pi_\beta}$ we get, $\delta = \frac{4\tau\pi_\beta}{\pi_\alpha}$. Therefore, the lower tail probability is bounded as follows:

$$\begin{aligned}
\mathsf{Pr}\left[X_\alpha - \frac{c\pi_\alpha}{\pi_\beta} \leq -4\tau c | X_\beta = c\right] &< \exp\left(\frac{-c\pi_\alpha}{2\pi_\beta}\cdot\left(\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^2\right) \\
&= \exp\left(-8c\tau^2\frac{\pi_\beta}{\pi_\alpha}\right) \\
&= \mathcal{L}^{ovif}(\beta).
\end{aligned}$$

http://dx.doi.org/10.21622/ACE.2021.01.1.007

Now consider the upper tail probability

$$\Pr\left[X_\alpha - \frac{c\pi_\alpha}{\pi_\beta} \geq 4\tau c | X_\beta = c\right] \quad = \quad \Pr\left[X_\alpha \geq \frac{c\pi_\alpha}{\pi_\beta} + 4\tau c | X_\beta = c\right].$$

Using Chernoff's upper tail bounds as described in the proof of Lemma 1 and $\delta = 4\tau\pi_\beta/\pi_\alpha$, the upper tail probability is bounded as:

$$\Pr\left[X_\alpha - \frac{c\pi_\alpha}{\pi_\beta} \geq 4\tau c | X_\beta = c\right] < \left(\frac{\mathrm{e}^{\frac{4\tau\pi_\beta}{\pi_\alpha}}}{\left(1+\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^{1+\frac{4\tau\pi_\beta}{\pi_\alpha}}}\right)^{\frac{c\pi_\alpha}{\pi_\beta}} = \mathcal{U}^{ovif}(\beta).$$

Combining the two tail probabilities proves the Lemma. □

We assume the existence of a maximum transition probability among all probabilities associated with transitions to or from any given state in Assumption 1.

Assumption 1. *Consider an order-2 DBC $\mathcal{DC}$ that generates sequence H. Let $\hat{\mathcal{DC}}$ be the DBC reconstructed from H. Given a state $\beta \in \Sigma^w_{\mathsf{DNA}}$ in the DBC $\mathcal{DC}$, define trans(β) as the set of all transitions of the form $\beta \to \beta[2\cdots w]\sigma$ or $\sigma\beta[1\cdots w-1] \to \beta$, for $\sigma \in \Sigma_{\mathsf{DNA}}$. For a positive constant s, $s > 0$, there exists a maximum transition $t^* \in$ trans(β) in $\mathcal{DC}$ such that, for all $t \in$ trans$(\beta) \setminus \{t^*\}$, $p(t^*) - p(t) > s$, where $p(t)$ denotes the probability associated with the transition t. For $0 \leq \varsigma \leq 1$, The probability that the same transition t^* is also the maximum probability transition for state β in $\hat{\mathcal{DC}}$ is given by*

$$\Pr\left[p(t^*) - p(t) > s\right] = \varsigma.$$

Given $\beta \in \Sigma^w_{\mathsf{DNA}}$, we define the maximum β-transition t^*_β as the transition in trans(β) having maximum frequency. The frequency of t^*_β is the *vertex isolation frequency* of β. Define $\mathcal{S}(\beta)$ as the state at which t^*_β starts and $\mathcal{E}(\beta)$ as the state at which t^*_β ends. Define $\mathcal{T}(\beta)$ as the label of t^*_β. When t^*_β is directed away from β, $\mathcal{S}(\beta) = \beta$, $\mathcal{E}(\beta) = \beta[2\ldots w]\sigma$, and $\mathcal{T}(\beta) = \beta\sigma$, for some $\sigma \in \Sigma_{\mathsf{DNA}}$. When t^*_β is directed into β, $\mathcal{S}(\beta) = \sigma\beta[1\ldots w-1]$, $\mathcal{E}(\beta) = \beta$, and $\mathcal{T}(\beta) = \sigma\beta$, for some $\sigma \in \Sigma_{\mathsf{DNA}}$.

The L_1 distance between the θ_2^{ovif} signatures of sequences generated by the same DBC is bounded in Theorem 2.

Theorem 2. *Let H_1 and H_2 be two genomic sequences generated by the same DBC of order 2. Let θ_1^{ovif} and θ_2^{ovif} be their respective order-2 θ^{ovif} signatures. Then for any $\tau > 0$,*

$$\Pr\left[d\left(\frac{\theta_1^{ovif}}{4}, \frac{\theta_2^{ovif}}{4}\right) > 32\tau\right] \quad < \quad 2\varsigma^2 \sum_{\beta\in\mathcal{S}^2} \left(\mathcal{L}^{ovif}(\beta) + \mathcal{U}^{ovif}(\beta)\right).$$

Proof. Using the results from Lemma 2, Assumption 1, and the union bound we get

$$\Pr\left[\left|\frac{\theta_1^{ovif}}{4} - \mathsf{E}\left[\frac{\theta_1^{ovif}}{4}\right]\right| > 16\tau\right] \quad < \quad \varsigma^2 \sum_{\beta\in\mathcal{S}^2} \left(\mathcal{L}^{ovif}(\beta) + \mathcal{U}^{ovif}(\beta)\right) \text{ and}$$

$$\Pr\left[\left|\frac{\theta_2^{ovif}}{4} - \mathsf{E}\left[\frac{\theta_2^{ovif}}{4}\right]\right| > 16\tau\right] \quad < \quad \varsigma^2 \sum_{\beta\in\mathcal{S}^2} \left(\mathcal{L}^{ovif}(\beta) + \mathcal{U}^{ovif}(\beta).\right)$$

The component-wise expected values in $\theta_1^{ovif}/4$ and $\theta_2^{ovif}/4$ are the same. Therefore,

$$d\left(\frac{\theta_1^{ovif}}{4}, \frac{\theta_2^{ovif}}{4}\right) = d\left(\frac{\theta_1^{ovif}}{4} - \mathsf{E}\left[\frac{\theta_1^{ovif}}{4}\right], \frac{\theta_2^{ovif}}{4} - \mathsf{E}\left[\frac{\theta_2^{ovif}}{4}\right]\right).$$

http://dx.doi.org/10.21622/ACE.2021.01.1.007

The theorem follows from the following:

$$\begin{aligned}\Pr\left[d\left(\frac{\theta_1^{ovif}}{4},\frac{\theta_2^{ovif}}{4}\right)>32\tau\right] &\leq \Pr\left[d\left(\frac{\theta_1^{ovif}}{4},\mathsf{E}\left[\frac{\theta_1^{ovif}}{4}\right]\right)>16\tau\right]+\\ &\quad \Pr\left[d\left(\frac{\theta_2^{ovif}}{4},\mathsf{E}\left[\frac{\theta_2^{ovif}}{4}\right]\right)>16\tau\right]\\ &< 2\varsigma^2\sum_{\beta\in\mathcal{S}^2}\left(\mathcal{L}^{ovif}(\beta)+\mathcal{U}^{ovif}(\beta)\right).\end{aligned}$$

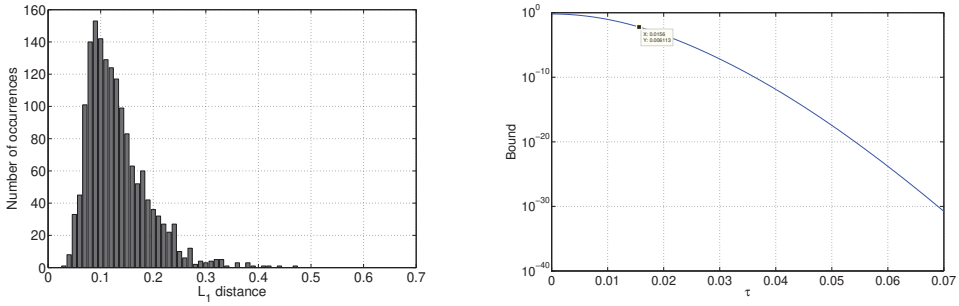
We now analyze the behavior of the upper bound in Theorem 2 when applied to real data. For a randomly selected set of genomic sequences, 100 pairs of sequences of length 100 kb each were randomly and independently sampled from each genomic sequence. For each pair, their $\theta_2^{ovif}/4$ signatures were computed and the L_1 distance between them was noted. Figure 14(a) plots the distribution of these distances. Note that a distance greater than approximately 0.5 marks a large and unlikely separation. The τ value corresponding to a distance of 0.5 is $0.5/32 = 0.0156$, whose corresponding upper bound of probability is very low as observed in Figure 14(b).

Next, we combine the properties of the $\hat{\pi}_2$ and $\theta_2^{ovif}/4$ signatures to derive the separation between θ_2^{dbc} signatures of sequences generated by the same DBC.

Separation between θ_2^{dbc} signatures derived from sequences generated by the same DBC

For sequences hypothesized to be generated by the same de Bruijn chain, Theorem 3 proves that the separation between their θ_w^{dbc} signatures is less than a specified threshold with high probability.

Theorem 3. *Let $\mathcal{DC}$ be an order s DBC. Let H_1 and H_2 be two genomic sequences of length n generated independently by $\mathcal{DC}$. Let θ_1^{dbc} and θ_2^{dbc} be their respective order-w*



(a) (b)

Fig 14. (a) Plot of distribution of L_1 distances between $\theta_2^{ovif}/4$ signatures of sequences from the same genome. τ can be derived by dividing each L_1 distance by 32. Note that approximately 0.5 distance or $\tau = 0.0156$ indicates a large and unlikely separation between two $\theta_2^{ovif}/4$ signatures. (b) Plot of upper bounds of separation between $\theta_2^{ovif}/4$ signatures of sequences from the same DBC using the theoretical expression derived in Theorem 2. Note that the probability for $\tau > 0.0156$ is 0.006113, which is low. $n = 10000$, $\varsigma = 0.75$, and a uniform stationary distribution were used for computing the bounds in (b).

DBC signatures. Similarly, let $\hat{\pi}^1$ and $\hat{\pi}^2$ be their order-2 stationary distributions and θ_1^{ovif} and θ_2^{ovif} be their order-2 OVIF signatures, respectively. Then,

$$\Pr\left[d\left(\theta_1^{dbc},\theta_2^{dbc}\right)>64\tau\right] < 2\cdot\sum_{\beta\in\mathcal{S}^2}\left(\mathcal{L}^{\pi}(\beta)+\mathcal{U}^{\pi}(\beta)\right)+2\varsigma^2\sum_{\beta\in\mathcal{S}^2}\left(\mathcal{L}^{ovif}(\beta)+\mathcal{U}^{ovif}(\beta)\right).$$

Proof. Note that $\theta_2^{dbc} = \hat{\pi}^2\cdot\theta_2^{ovif}/4$. Using the union bound we have

$$\Pr\left[d\left(\theta_1^{dbc},\theta_2^{dbc}\right)>64\tau\right] \leq \Pr\left[d\left(\hat{\pi}^1,\hat{\pi}^2\right)>32\tau\right]+\Pr\left[d\left(\theta_1^{ovif},\theta_2^{ovif}\right)>32\tau\right].$$

The theorem follows using the results from Theorems 1 and 2. □

http://dx.doi.org/10.21622/ACE.2021.01.1.007

Separation between θ_2^{dbc} signatures of sequences generated by different DBCs

Let H_1 and H_2 be genomic sequences of length n, generated independently by two different order-2 DBCs $\mathcal{DC}_1$ and $\mathcal{DC}_2$, respectively. Let θ_1^{dbc} and θ_2^{dbc} be their order-w DBC signatures. Let $\hat{\pi}^1$ and $\hat{\pi}^2$ be their estimated stationary distributions and θ_1^{ovif} and θ_2^{ovif} be their OVIF signatures.

Then, the distance $d\left(\theta_1^{dbc}, \theta_2^{dbc}\right)$ can distinguish $\mathcal{DC}_1$ and $\mathcal{DC}_2$. Assumptions 2 formalizes the separation of estimated stationary distributions of genomic sequences obtained from different organisms, while Assumption 3 formalizes the probability of the maximum transition being different for a given state using genomic sequences obtained from different organisms.

Assumption 2. $d\left(\mathsf{E}\left[\hat{\pi}^1\right], \mathsf{E}\left[\hat{\pi}^2\right]\right) > 3 \cdot 16\tau.$

Assumption 3. $d\left(\mathsf{E}\left[\theta_1^{ovif}\right], \mathsf{E}\left[\theta_2^{ovif}\right]\right) > 3 \cdot 16\tau.$

Theorem 4. *If there exist constants γ and ν as in Assumptions 2 and 3, then*

$$\begin{aligned}\Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2 \cdot 16\tau\right] \quad \geq \quad & 1 - \Pr\left[d\left(\theta_1^{dbc}, \mathsf{E}\left[\theta_1^{dbc}\right]\right) \geq 2 \cdot 16\tau\right] - \\ & \Pr\left[d\left(\theta_2^{dbc}, \mathsf{E}\left[\theta_2^{dbc}\right]\right) \geq 2 \cdot 16\tau\right].\end{aligned}$$

Proof. Treating $d\left(\theta_1^{dbc}, \theta_2^{dbc}\right)$, $d\left(\theta_1^{dbc}, \mathsf{E}\left[\theta_1^{dbc}\right]\right)$, $d\left(\theta_2^{dbc}, \mathsf{E}\left[\theta_2^{dbc}\right]\right)$, and $d\left(\mathsf{E}\left[\theta_1^{dbc}\right], \mathsf{E}\left[\theta_2^{dbc}\right]\right)$ as distances d, d_1, d_2, and d_3, respectively, in 1-dimensional space we obtain,

$$\begin{aligned} d_3 \quad &\leq \quad d + d_1 + d_2 \\ \Pr\left[d_3 \geq 6 \cdot 16\tau\right] \quad &\leq \quad \Pr\left[d \geq 2 \cdot 16\tau\right] + \Pr\left[d_1 \geq 2 \cdot 16\tau\right] + \Pr\left[d_2 \geq 2 \cdot 16\tau\right]. \end{aligned}$$

From Assumptions 2 and 3 we obtain, $d\left(\mathsf{E}\left[\theta_1^{dbc}\right], \mathsf{E}\left[\theta_2^{dbc}\right]\right) \geq 6 \cdot 16\tau$. We have

$$\begin{aligned} 1 \quad &\leq \quad \Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2 \cdot 16\tau\right] + \\ &\qquad \Pr\left[d\left(\theta_1^{dbc}, \mathsf{E}\left[\theta_1^{dbc}\right]\right) \geq 2 \cdot 16\tau\right] + \\ &\qquad \Pr\left[d\left(\theta_2^{dbc}, \mathsf{E}\left[\theta_2^{dbc}\right]\right) \geq 2 \cdot 16\tau\right] \\ \Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2 \cdot 16\tau\right] \quad &\geq \quad 1 - \Pr\left[d\left(\theta_1^{dbc}, \mathsf{E}\left[\theta_1^{dbc}\right]\right) \geq 2 \cdot 16\tau\right] - \\ &\qquad \Pr\left[d\left(\theta_2^{dbc}, \mathsf{E}\left[\theta_2^{dbc}\right]\right) \geq 2 \cdot 16\tau\right]. \end{aligned}$$

The theorem follows. □

We demonstrate Assumptions 2 and 3 using sequences from the species *C. elegans* and *P. falciparum*. Figure 15 presents the distribution of L_1 distances between θ_2^{dbc} signatures of pairs of 10 kilobase long sequences randomly sampled from the above two species, respectively. The actual distance between the expected values of $\hat{\pi}^1$ and $\hat{\pi}^2$ is 0.4735. From Assumption 2, we have $\tau < 0.4735/48 = 0.0099$. Using $d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2 \cdot 16\tau$ gives $d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 32\tau$. For $\tau < 0.0099$, $32\tau < 0.3168$, and the probability $\Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2 \cdot 16\tau\right]$ is large as seen in Figure 15. A similar scenario is observed for Assumption 3. The L_1 distance between the expected values of the θ_1^{ovif} and θ_2^{ovif} 0.374622, which leads to τ being less than $0.374622/48 = 0.0078$. For these values of τ, the probability $\Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2 \cdot 16\tau\right]$ is high.

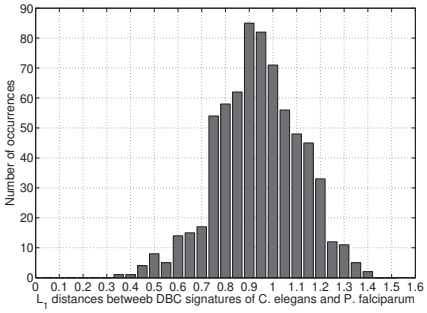


Fig 15. Distribution of L_1 distances between θ_2^{dbc} signatures of pairs of 10 kilobase long sequences randomly sampled from the two species C. elegans and P. falciparum.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

In Theorem 4, each negative term in the right hand side is very small, making the total probability on the right hand side a very large value. Theorem 4 states that the probability that the separation between the θ_w^{dbc}s of two sequences hypothesized to be generated by different DBCs exceeds a given threshold is very high.

Time complexity

The θ_2^{dbc} for a sequence of length n can be computed in $O(n + 16 \log 16 + 4096)$ time and space. In general, the complexity of computing the order-w θ_w^{dbc} signature for a sequence of length n is $O(n + 4^w \log 4^w + (4^w)^3)$. The $(4^w)^3$ factor is contributed by the Cholesky decomposition performed by MATLAB to compute the stationary distribution. For small $w \in [1, 4]$, we observed that the time complexity was dominated by n, as we would expect.

Conclusions

We have examined genomic signatures from the point of view of accurate identification of the origin of short unknown DNA sequences. The genomic signatures introduced in this paper are derived from the structure and properties of de Bruijn chains. When a sample sequence is sufficiently long, the target organism for the sample can be retrieved by querying a database of signatures. Given an unknown DNA sequence, its possible high-level location in the phylogenetic tree can be predicted using the combination of the θ_2^{dbc} and θ^{dor} signatures, after which its origin and closest relatives can be predicted using the θ^{dbc} signature alone.

We have demonstrated both theoretically and empirically that θ^{dbc} is a powerful signature, able to efficiently identify the origin of an unknown genomic sequence as short as a few kilobases. This implies that the origin and the closest relatives of an unknown sequence can be identified with very little actual sequencing. We also observed the effect of order on efficiency of the θ^{dbc} signature. In continuing work, we are exploring the effect of size of the signature database on short sequence target prediction efficiency. We are also studying the phylogeny implied by distances between θ^{dbc} signatures and the extent to which this phylogenetic structure is conserved on random sampling of short sequences for phylogenetic reconstruction.

Acknowledgments

This work was supported by the United States National Science Foundation Grant NSF ITR 0428344. We thank Naren Ramakrishnan, João Setubal, Richard Helm, and Anil Shende for invaluable suggestions. We also thank the two anonymous reviewers for their careful feedback.

References

1. Kaposi-Novak P, Lee JS, Gomez-Quiroz L, Coulouarn C, Factor VM, Thorgeirsson SS. Met-regulated expression signature defines a subset of human hepatocellular carcinomas with poor prognosis and aggressive phenotype. Journal of Clinical Investigation. 2006;116:1582–1595.

2. Mandruzzato S, Callegaro A, Turcatel G, Francescato S, Montesco MC, Chiarion-Sileni V, et al. A gene expression signature associated with survival in metastatic melanoma. Journal of Translational Medicine. 2006;4(50). doi:10.1186/1479-5876-4-50.

3. Anguiano A, Potti A. Genomic signatures individualize therapeutic decisions in non-small-cell lung cancer. Future Drugs. 2007;7(6):837–844.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

4. Dressman HK, Bild A, Garst J, Jr DH, Potti A. Genomic signatures in non-small-cell lung cancer: Targeting the targeted therapies. Current Oncology Reports. 2006;8(4):252–257.

5. Urquidia V, Goodison S. Genomic signatures of breast cancer metastasis. Cytogenetic and Genome Research. 2007;118:116–129.

6. Mendiratta P, Febbo PG. Genomic Signatures Associated with the Development, Progression, and Outcome of Prostate Cancer. Molecular Diagnosis and Therapy. 2007;11(6):345–354.

7. Chang JT, Nevins JR. GATHER: A systems approach to interpreting genomic signatures. Bioinformatics. 2006;22(23):2926–2933.

8. Normark BB, Judson OP, Moran NA. Genomic signatures of ancient asexual lineages. Biological Journal of the Linnean Society. 2003;79:69–84.

9. Cannon CH, Kua CS, Lobenhofer EK, Hurban P. Capturing genomic signatures of DNA sequence variation using a standard anonymous microarray platform. Nucleic Acids Research. 2006;34(18). doi:10.1093/nar/gkl478.

10. Hande MP, Azizova TV, Geard CR, Burak LE, Mitchell CR, Khokhryakov VF, et al. Past Exposure to Densely Ionizing Radiation Leaves a Unique Permanent Signature in the Genome. The American Society of Human Genetics. 2003;72:1162–1170.

11. Mao JH, Li JZ, Jiang T, Li Q, Wu D, Perez-Losada J, et al. Genomic Instability in Radiation-Induced Mouse Lymphoma from P53 Heterozygous Mice. Oncogene. 2005;24(53):7924–7934.

12. Maddison DR, Schulz KS, Maddison WP. The Tree of Life Web Project. Zootaxa. 2007;1668:19–40.

13. Carbone A, Kepes F, Zinovyev A. Codon Bias Signatures, Organization of Micro-organisms in Codon Space, and Lifestyle. Molecular Biology and Evolution. 2005;22(3):547–561.

14. Coenye T, Vandamme P. Use of the Genomic Signature in Bacterial Classification and Identification. Systematic and Applied Microbiology. 2004;27(2):175–185.

15. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences. Molecular Biology and Evolution. 1999;16(10):1391–1399.

16. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. Detection and Characterization of Horizontal Transfers in Prokaryotes Using Genomic Signature. Nucleic Acids Research. 2005;33(1):12 pages.

17. Fertil B, Massin M, Lespinats S, Devic C, Dumee P, Giron A. GENSTYLE: Exploration and analysis of DNA sequences with genomic signature. Nucleic Acids Research. 2005;33 (Web Server issue):W512–W515.

18. Jernigan RW, Baran RH. Pervasive Properties of the Genomic Signature. BMC Genomics. 2002;3:9 pages.

19. Karlin S, Burge C. Dinucleotide Relative Abundance Extremes — A Genomic Signature. Trends in Genetics. 1995;11(7):283–290.

20. Karlin S, Mrazek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology. 1997;179(12):3899–3913.

21. Sandberg R, Branden CI, Ernberg I, Coster J. Quantifying the Species-Specificity in Genomics Signatures, Synonymous Codon Choice, Amino Acid Usage, and G+C Content. Gene. 2003;311:35–42.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

22. Teeling H, Meyerdierks A, Buaer M, Amann R, Glockner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environmental Microbiology. 2004;6:938–947.

23. van Passel MWJ, Bart A, Thygesen HH, Luyf ACM, van Kampen AHC, van der Ende A. An Acquisition Account of Genomic Islands Based on Genome Signature Comparisons. BMC Genomics. 2005;6:10 pages.

24. van Passel MWJ, Kuramae EE, Luyf ACM, Bart A, Boekhout T. The reach of the genome signature in prokaryotes. BMC Evolutionary Biology. 2006;6(84):8 pages. doi:10.1186/1471-2148-6-84.

25. Dalevi D, Dubhashi D, Hermansson M. Bayesian Classifiers for Detecting HGT Using Fixed and Variable Order Markov Models of Genomic Signatures. Bioinformatics. 2006;22(5):517–522.

26. Campbell AM, Mrazek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States Of America. 1999;96:9184–9189.

27. Gentles AJ, Karlin S. Genome-Scale Compositional Comparisons in Eukaryotes. Genome Research. 2001;11:540–546.

28. Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Research. 1990;18(8):2163–2170.

29. Dutta C, Das J. Mathematical characterization of Chaos Game Representation: New algorithms for nucleotide sequence analysis. Journal of Molecular Biology. 1992;228(3):715–719.

30. Hill KA, Schisler NJ, Singh SM. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. Journal of Molecular Evolution. 1992;35(3):261–269.

31. Pevzner PA. DNA Physical Mapping and Alternating Eulerian Cycles in Colored Graphs. Algorithmica. 1995;13(1-2):77–105.

32. Pevzner PA, Tang HX, Waterman MS. An Eulerian Path Approach to DNA Fragment Assembly. Proceedings of The National Academy of Sciences of the United States Of America. 2001;98(17):9748–9753.

33. Zhang Y, Waterman MS. An Eulerian Path Approach to Global Multiple Alignment for DNA Sequences. Journal of Computational Biology. 2003;10(6):803–819.

34. Raphael B, Zhi DG, Tang HX, Pevzner P. A Novel Method for Multiple Alignment of Sequences with Repeated and Shuffled Elements. Genome Research. 2004;14(11):2336–2346.

35. Zhang Y, Waterman MS. An Eulerian Path Approach to Local Multiple Alignment for DNA Sequences. Proceedings of The National Academy of Sciences of the United States Of America. 2005;102(5):1285–1290.

36. Heath LS, Pati A. Genomic signatures from DNA word graphs. In: Proceedings of the Third International Symposium on Bioinformatics Research and Applications (ISBRA). Springer-Verlag; 2007. p. 317–328.

37. Heath LS, Pati A. Genomic signatures in de Bruijn chains. In: Proceedings of the Workshop on Algorithms in Bioinformatics (WABI). Springer-Verlag; 2007. p. 216–227.

38. Heath LS, Pati A. Predicting Markov chain order in genomic sequences. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE Computer Society; 2007. p. 159–164.

39. Pati A, Heath LS, Kyrpides NC, Ivanova N. ClaMS: A Classifier for Metagenomic Sequences. Standards in Genomic Sciences. 2011;5(2):248–253.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

40. Pati A. Graph-based Genomic Signatures [PhD Dissertation]. Virginia Tech. Blacksburg, Virginia; 2008.

41. Konstantinidis KT, Tiedje JM. Towards a Genome-based Taxonomy for Prokaryotes. Journal of Bacteriology. 2005;187(18):6258–6264.

42. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA Hybridization Values and Their Relationship to Whole-Genome Sequence Similarities. International Journal of Systematic and Evolutionary Microbiology. 2007;57:81–91.

43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. Journal of Molecular Biology. 1990;215(3):403–410.

44. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and Open Software for Comparing Large Genomes. Genome Biology. 2004;5(2):9 pages.

45. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and Taxonomy in Diagnostics for Food Security: Soft-Rotting Enterobacterial Plant Pathogens. Analytical Methods. 2016;8(1):12–24.

46. Marakeby H, Badr E, Torkey H, Song Y, Leman S, Monteil CL, et al. A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature. PLOS One. 2014;9(2):12 pages.

47. Tian L, Huang CJ, Mazloom R, Heath LS, Vinatzer BA. Linbase: A Web Server for Genome-Based Identification of Prokaryotes as Members of Crowdsourced Taxa. Nucleic Acids Research. 2020;48(W1):W529–W537.

48. Tian L, Mazloom R, Heath LS, Vinatzer BA. Linflow: A Computational Pipeline That Combines an Alignment-Free with an Alignment-Based Method to Accelerate Generation of Similarity Matrices for Prokaryotic Genomes. Peerj. 2021;9:17 pages.

49. Vinatzer BA, Elmarakeby HA, Weisberg AJ, Monteil CL, Heath LS. A New Exclusively Genome-Based Species-Independent Taxonomic Framework for All Life Forms Applied to Pseudomonas syringae. Phytopathology. 2015;105(11):143.

50. Vinatzer BA, Tian L, Heath LS. A Proposal for a Portal to Make Earth'S Microbial Diversity Easily Accessible and Searchable. Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology. 2017;110(10):1271–1279.

51. Weisberg AJ, Elmarakeby HA, Heath LS, Vinatzer BA. Similarity-Based Codes Sequentially Assigned to Ebolavirus Genomes Are Informative of Species Membership, Associated Outbreaks, and Transmission Chains. Open Forum Infectious Diseases. 2015;2(1):11 pages.

52. Broder AZ. On the Resemblance and Containment of Documents. In: Compression and Complexity of Sequences 1997 — Proceedings; 1998. p. 21–29.

53. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: Fast Genome and Metagenome Distance Estimation Using MinHash. Genome Biology. 2016;17:14 pages.

54. Baker DN, Langmead B. Dashing: Fast and Accurate Genomic Distances with HyperLogLog. Genome Biology. 2019;20(1):12 pages.

55. Brown CT, Irber L. sourmash: a library for MinHash sketching of DNA. The Journal of Open Source Software. 2016;1:27.

56. Marcais G, DeBlasio D, Pandey P, Kingsford C. Locality-Sensitive Hashing for the Edit Distance. Bioinformatics. 2019;35(14):I127–I135.

http://dx.doi.org/10.21622/ACE.2021.01.1.007

57. Zhao XF. BinDash, Software for Fast Genome Distance Estimation on a Typical Personal Laptop. Bioinformatics. 2019;35(4):671–673.

58. Rosenberg AL, Heath LS. Graph Separators, With Applications. Frontiers of Computer Science. Kluwer Academic/Plenum Publishers; 2000.

59. Fickett JW, Torney DC, Wolf DR. Base Compositional Structure of Genomes. Genomics. 1992;13(4):1056–1064.

60. Feller W. An Introduction to Probability Theory and Its Applications. vol. I. 3rd ed. New York: John Wiley & Sons Inc.; 1968.

61. Solan E, Vieille N. Perturbed Markov chains. Journal of Applied Probability. 2003;40:107–122.

62. Mitzenmacher M, Upfal E. Probability and Computing. 1st ed. Cambridge University Press; 2005.