Graph-Based Genomic Signatures

Amrita Pati

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Lenwood S. Heath, Chair
Richard F. Helm
Narendran Ramakrishnan
João Carlos Setubal
Anil M. Shende

April 14, 2008
Blacksburg, Virginia

Keywords: Genomes, Genomic signatures, de Bruijn graphs, Markov chains

Graph-based genomic signatures

Amrita Pati

(ABSTRACT)

Genomes have both deterministic and random aspects, with the underlying DNA sequences exhibiting 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. This work examines the unique manner in which oligonucleotides fit together to comprise a genome, within a graph-theoretic setting. A *de Bruijn chain (DBC)* is a marriage of a de Bruijn graph and a finite Markov chain. By representing a DNA sequence as a walk over a DBC and retaining specific information at nodes and edges, we are able to obtain the de Bruijn chain genomic signature (DBCGS), based on both graph structure and the stationary distribution of the DBC. We demonstrate that DBCGS is information-rich, efficient, sufficiently representative of the sequence from which it is derived, and superior to existing genomic signatures such as the dinucleotides odds ratio and word frequency based signatures. We develop a mathematical framework to elucidate the power of the DBCGS signature to distinguish between sequences hypothesized to be generated by DBCs of distinct parameters. We study the effect of order of the DBCGS signature on accuracy while presenting relationships with genome size and genome variety. We illustrate its practical value in distinguishing genomic sequences and predicting the origin of short DNA sequences of unknown origin, while highlighting its superior performance compared to existing genomic signatures including the dinucleotides odds ratio.

Additionally, we describe details of the CMGS database, a centralized repository for raw and value-added data particular to *C. elegans*.

This work was supported by NSF-ITR Grant-0428344 for the Computational Models for Gene Silencing project.

Dedication

To Mummy and Baba, for introducing to me the scientific method and instilling in me the belief that all dreams can be realized

To Juju, for always placing me first

To Swaroop, for being my inspiration for perfection

Acknowledgments

I thank Dr. Lenny Heath for his invaluable guidance, advise, patience, encouragement, and his confidence in me. Having him as my advisor has in many good ways enriched my time as a Ph.D. candidate and shaped my approach towards research. I thank my committee members: Dr. Naren Ramakrishnan, Dr. João Setubal, Dr. Dr. Richard Helm, and Dr. Anil Shende for their useful suggestions, directions, and help. My friends in Torgersen Hall were always there with interesting conversations to brighten up dull times. Many thanks to Allan, Douglas, Jon, Ying, and others for their lively presence.

My research work has been supported by the National Science Foundation's NSF-ITR Grant-0428344 for the Computational Models for Gene Silencing project and I am grateful for the support. I thank the Department of Computer Science at Virginia Tech for the teaching assistantship that I have received and for making graduate school a pleasant experience. Rob, our system administrator, has been very cooperative with every request and I thank him for his help and many interesting discussions.

I thank my roommate Vidya for being there at all times, and the Shendes for many interesting times and a lot of help. Friends in my music and dance groups were responsible for many fun times.

I thank my parents and my brother Animesh for their confidence, patience, and love. I also thank Aaee, Aja, JJMa, and JJBapa for their blessings, and my uncles and aunts for their love and support.

I thank Radha, Ravi, Uncle, and Aunty for their love and support, and Rahul and Malavika for being adorable.

I thank Swaroop for his love, support, inspiration, and for being a part of my life.

Finally, I thank God for making everything possible.

Attribution

Several colleagues and coworkers aided in the writing and research behind Appendix A. A brief description of their background and their contributions are included here.

Prof. Lenwood S. Heath - Ph.D. (Department of Computer Science, Virginia Tech) is the primary advisor. Prof. Heath provided many useful insights during the construction of the database.

Prof. Naren Ramakrishnan - Ph.D. (Department of Computer Science, Virginia Tech) advised on aspects of the data mining engine integrated with CMGSDB.

Ying Jin - Graduate student (Department of Computer Science, Virginia Tech) implemented the data-mining engine.

Prof. Richard F. Helm - Ph.D. (Department of Biochemistry, Virginia Tech) advised on the integration of RNAi phenotypes from multiple sources.

Karsten Klage - Ph.D. (Department of Biochemistry, Virginia Tech) helped in the classification of RNAi phenotypes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

The last 25 years have seen tremendous progress in our understanding of biological paradigms. The sequencing of genomes has opened up aspects of genomic sequences that had never been envisaged. Since the first sequencing of a genome (the 5386-base long bacteriophage ϕ-X174) in 1977, several genomes have been sequenced, including the 3.2×10^9 bases long human genome, whose sequencing was completed in 2003. A comprehensive source of detailed information regarding complete and ongoing genome projects around the world is the Genomes Online Database [79], which housed approximately 3250 incomplete and approximately 800 completed genome projects as of January 2008. Of these, approximately 2300 are bacterial genome projects, approximately 150 are archaeal genome projects, and approximately 1000 are eukaryotic genome projects.

Analysis of such high-throughput genomic data has necessitated the application of computational models and techniques. An organism's genomic sequence encodes all of its individual traits at various scales ranging from individual nucleotide (A, C, G, T) compositions to gene orders in large genomic regions. 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 species to another, it can serve as a unique mathematical structure, a string, representing a species. 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 $\theta(H)$, typically a vector of numbers 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 even for relatively short lengths of H. The intent is that the signatures of other large substrings from $\mathcal{G}$ be highly similar to $\theta(H)$ and distinguishably different 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 from relatively short sequences.

The term "genomic signature" must not be confused with the term "gene expression signature" [67, 84], although the two terms have been used interchangeably in a number of works [99, 7, 32, 124, 85, 19]. 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 [67, 84]. Normark et al. [93] 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. [17] have used it to represent probe sequences that are short (25 bases and less) primers that are hyper-dispersed in a probability space of sequences and generated without 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 intra-chromosomal aberrations [50, 65]. In this work, a genomic signature, as defined in the previous paragraph, is a unique mathematical structure strictly computed from sequence data and conserved across reasonably large (a few kilobases) subsequences of a genome for a wide range of subsequence lengths.

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 (Chapter 2). We derive the θ^{dbc} signature by thinking of a genomic sequence as a walk over a suitably defined DBC. We then combine characteristic properties of the stationary distribution of the underlying Markov chain with the manner in which the DBC disintegrates on deleting edges in a systematic manner, to obtain 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 (See Chapter 4). In this work, 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 signature $\theta_w(X)$ of a pre-defined type at order w, is computed from X and compared to the

same signature at the same order w for the genomic sequences of all species with sequenced genomes using an algorithm proposed in this work. 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 di-nucleotides 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.

Several applications of genomic signatures are possible, some of which 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 [81]. 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. The application of genomic signatures to binning metagenomic data can also be perceived.

Another aspect of sequences that are hypothesized to be generated by a Markov chain $\mathcal{M}$ is the order of the underlying Markov chain $\mathcal{M}$. We hypothesize that each genome is generated, within a reasonable approximation, by a Markov chain of unknown order. Given a sequence H, we call such a Markov chain $\mathcal{M}$ that generates H, the *generating Markov chain* of H. Estimating the order of the generating Markov chain will assist in understanding biological phenomena such as a difference in frequencies of observed DNA word[1] patterns and repeats in the genome [52, 54]. We present an algorithm that uses frequencies of DNA words at various orders to estimate the order of the Markov chain that generates a given sequence. While existing methods are based on principles of entropy estimation, maximal fluctuation, and maximum likelihood estimation, in this work, we propose a randomized algorithm for estimating the order of a generating Markov chain within a framework of probability distributions of its states and transitions.

This dissertation is organized as follows. Chapter 2 defines the fundamental mathematical concepts used in this work. It also lays down the basic conceptual framework within which the rest of this work is organized, and establishes notation. Chapter 3 precisely defines the computational problem at hand and describes its various sub-problems and derivatives that are addressed in this work. In Chapter 4 we describe relevant research in the scientific literature related to this work and highlight their key contributions and important results. Chapter 5 introduces and defines purely graph-based signatures. It establishes graph-based

[1] A *DNA word* of length w is a string in Σ_{DNA}^{w}.

signatures as discriminating between species and conserved within a species. We also present comparisons between existing signatures in the literature and purely graph-based signatures. In Chapter 6 we introduce and define the novel de Bruijn chain signature θ^{dbc}. We explore its properties, both within a theoretical framework, and using experimental methods. We establish its accuracy in origin prediction of unknown DNA sequences as greater than the accuracy of any existing methods and present relevant results. In this chapter, we establish the superiority of the θ^{dbc} signature over existing signatures. We study the variation in accuracy of origin prediction with varying sample sequence length as well as order of the signature. In Chapter 7, we explore the problem of predicting the order of the generating Markov chain of a given sequence. We examine existing methods for doing so, and present a novel, sampling-based approach to predict the order of the generating Markov chain of a given sequence. The complexity of this algorithm is much less than that of existing methods. The dissertation is concluded in Chapter 8, where we summarize our insights into the area of genomic signatures and discuss future directions of research.

The Appendix describes in detail the database for the Computational Models for Gene Silencing (CMGS) project. Although unrelated to the central theme of this dissertation, the grant for the CMGS project was responsible for funding this work.

Chapter 2

Definition, Notation, and Preliminaries

In this chapter, we define the essential concepts for a study of genomic sequences and genomic signatures within a graph-theoretic and formal language framework. We build the fundamental framework and establish necessary notation.

2.1 Formal language background

Hopcroft and Ullman [60] and Lewis and Papadimitriou [77] are standard references for formal language concepts. An *alphabet* Σ is a non-empty, finite set of symbols. In particular, the *DNA alphabet* is $\Sigma_{\mathrm{DNA}} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and the *binary alphabet* is $\Sigma_{\mathrm{B}} = \{0, 1\}$. A *string* over Σ is a finite sequence of symbols, written as a concatenation of symbols. For a string u over Σ, the *length* $|u|$ of u is its length as a sequence. The sequence ATGCCA is a length-6 string over Σ_{DNA}. The *empty string* λ is the unique string of length 0.

A single chromosome in a genome is typically written as the string 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.

The set of all strings over Σ is Σ^*, an infinite set. If $u, v \in \Sigma^*$, and $\cdot$ is the concatenation operator, the *concatenation* $u \cdot v$ of u and v is the sequence obtained by putting u before v; the concatenation $u \cdot v$ is typically abbreviated uv. Clearly, $|uv| = |u| + |v|$. The concatenation of the two strings GGAG and TCC

from Σ^*_{DNA} is GGAGTCC. The set Σ^* is a monoid [61] under concatenation, with identity element λ. Let $u = \sigma_1\sigma_2\cdots\sigma_n \in \Sigma^*$ be a string of length n. A string $v \in \Sigma^*$ is a *substring* of u if there exist strings $x, y \in \Sigma^*$ such that $u = xvy$. A string $v \in \Sigma^*$ is a *prefix* of u if there exists a string $x \in \Sigma^*$ such that $u = vx$. A string $v \in \Sigma^*$ is a *suffix* of u if there exists a string $x \in \Sigma^*$ such that $u = xv$. Including the empty string, u has $n+1$ prefixes and $n+1$ suffixes.

For strings x and y with $|x| \leq |y|$, $\text{occ}(x, y)$ is the count of occurrences of x as a substring of y. We indicate x being a substring of y by the expression $x \sqsubset y$. The *frequency* of x in y is

$$\text{freq}(x, y) = \frac{\text{occ}(x, y)}{|y| - |x| + 1}.$$

Fix a word length $w \geq 1$. The *order-w state space* is $\mathcal{S}^w = \Sigma^w_{\text{DNA}}$, the set consisting of the 4^w words of length w.

2.2 Markov chains

We are interested only in discrete-time stochastic processes that have a finite state space. Let $\mathcal{X} = \{X_i \mid 0 \leq i\}$ be a set of random variables, indexed by non-negative integers, over the same probability space, such that each X_i takes values in the *state space* $\mathcal{S}^w$ of $\mathcal{X}$. The set $\mathcal{X}$ has the *Markov property* if, for all $n \geq 0$ and all $0, 1, \ldots, n+1 \in \mathcal{S}^w$, we have

$$\Pr[X_{n+1} = n+1 \mid X_0 = 0, X_1 = 1, \ldots, X_n = n] \quad = \quad \Pr[X_{n+1} = n+1 \mid X_n = n].$$

If $\mathcal{X}$ has the Markov property, then it is a *discrete time, finite Markov chain*, or simply, a *Markov chain*. The Markov chain $\mathcal{M}$ is *homogeneous in time* if, for all $m, n \geq 0$ and all $j, k \in \mathcal{S}^w$, we have

$$\Pr[X_{m+1} = m+1 \mid X_m = m] \quad = \quad \Pr[X_{n+1} = m+1 \mid X_n = m].$$

We will assume all Markov chains are homogeneous in time. The resulting conditional probabilities

$$p_{kj} \quad = \quad \Pr[X_{n+1} = j \mid X_n = k]$$

are the *(stationary) transition probabilities* of $\mathcal{M}$.

Let $s = |\mathcal{S}^w|$. For ease of notation, we label the states in $\mathcal{S}^w$ so that $\mathcal{S}^w = \{1, 2, \ldots, s\}$. The *transition probability matrix of* $\mathcal{M}$ is the $s \times s$ matrix

$$P_M \quad = \quad \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,s} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s,1} & p_{s,2} & \cdots & p_{s,s} \end{pmatrix}$$

For all $n \geq 0$, let μ_n be the probability distribution on X_n. For a Markov chain $\mathcal{M}$, the choice of μ_0, the *initial probability distribution,* determines every other μ_n. In fact, we have that

$$\begin{pmatrix} \mu_n(1) & \mu_n(2) & \dots & \mu_n(s) \end{pmatrix} \quad = \quad \begin{pmatrix} \mu_0(1) & \mu_0(2) & \dots & \mu_0(s) \end{pmatrix} P_M^n,$$

written more simply as $\mu_n = \mu_0 P_M^n$. Let $i, j \in \mathcal{S}^w$ be fixed states of $\mathcal{M}$. State i is *recurrent* if

$$\Pr\left[X_n = i, \text{ for some } n \geq 1 \mid X_0 = i\right] \quad = \quad 1;$$

otherwise, state i is *transient.* State i is *absorbing* if $p_{i,i} = 1$. The *period* of state i is the greatest common divisor of the set

$$\{n \geq 1 \mid \Pr\left[X_n = i \mid X_0 = i\right] > 0\}.$$

State i is *periodic* if its period is greater than 1; otherwise, it is *aperiodic.* The state pair (i, j) *communicates* if

$$\Pr\left[X_n = j, \text{ for some } n \geq 1 \mid X_0 = i\right] \quad > \quad 0.$$

The Markov chain $\mathcal{M}$ is *ergodic* if every pair of states communicates and if every state is recurrent and aperiodic.

Let S be a genomic sequence of length n. The *generating Markov chain* of S, $\mathcal{G}(S)$, is defined as the yet to be characterized, hypothetical, Markov chain that generates S. Consider the transition probabilities computed for an order-w Markov chain using the counts of strings in $\mathcal{S}^w$ in S. Define the *empirical transition function* $P_{w,emp}(S)$ as the transition probability matrix obtained by enumerating all words of length w in S and calculating probabilities of transitions between them. Define the *derived transition function* $P_{w,der}(S)$ as the transition probability matrix obtained from $P_{w-1,emp}(S)$. Let $x, y \in \mathcal{S}^w$. Then,

$$P_{w,der}(x, y) \quad = \quad P_{w-1,emp}(x[2 \dots w-1], y[2 \dots w-1]),$$

and

$$P_{w,emp}(x, y) \quad = \quad \frac{\text{occ}\left(x \cdot y[w], S\right)}{\text{occ}\left(x, S\right)}.$$

Consider an empirical transition probability matrix $P_{3,emp}(s)$ over Σ_{DNA} and some string s.
Let $P_{3,emp}(CCG, CGT) = 0.053$. Then $P_{4,der}(ACCG, CCGT) = P_{3,emp}(CCG, CGT) = 0.053$.

A complex number λ that is a solution to the matrix equation $P_M v = \lambda v$ is an *eigenvalue* of P_M corresponding to the eigenvector v. As the matrix equation has s, not necessarily distinct, complex roots, we can list the s roots as $\lambda_1, \lambda_2, \dots, \lambda_s$ in decreasing order by modulus. The set $\{\lambda_1, \lambda_2, \dots, \lambda_s\}$ is the *spectrum* of P_M.

Let $x, y \in \mathcal{S}^w$. If π is a probability distribution on $\mathcal{S}^w$ such that

$$\sum_{x \in \mathcal{S}^w} \pi(x) P(x, y) \quad = \quad \pi(y),$$

then π is a *stationary distribution*. π is represented as an $n-$dimensional vector satisfying the property

$$\pi^T P \quad = \quad \pi^T.$$

The distribution at time step 0 is denoted by π_0. Additionally, the distribution at time step t is denoted by $\pi_t = \pi_0 P^t$. If $\pi_0 = \pi$, then $\pi_t = \pi_0 P^t = \pi$ for all t.

Two states $i, j \in \mathcal{S}^w$ in the Markov chain $\mathcal{M}$ *communicate* if state j can be reached from state i and state i can be reached from state j. Markov chain $\mathcal{M}$ is said to be *irreducible* when every pair of states in $\mathcal{S}^w$ communicate. The *period* of a state i is defined as the greatest common divisor g of the set

$$\{n \mid \Pr[X_n = i \mid X_0 = i] > 0\}.$$

A state is said to be *aperiodic* when $g = 1$.

An irreducible finite Markov chain is *ergodic* when all its states are aperiodic. For an *ergodic* Markov chain, π is unique and $\pi_t \to \pi$ as $t \to \infty$.

Further related material on Markov chains can be found in a number of references, including these [4, 5, 10, 14, 34, 46, 59, 104].

2.3 De Bruijn graphs

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 [103]. Figure 2.1 depicts de Bruijn graphs of orders 2 and 3 over the binary alphabet Σ_{B} and the de Bruijn graph of order 2 over the DNA alphabet Σ_{DNA}. As observed, the vertex sets of the binary de Bruijn graphs of orders 2 and 3 are the set of all binary strings of length 2 ($\{00, 01, 10, 11\}$) and the set of all binary strings of length 3 ($\{000, 001, 010, 011, 100, 101, 110, 111\}$), respectively. Similarly, the vertex set of the DNA de Bruijn graph of order 2 is

$$\Sigma^2_{\mathrm{DNA}} \quad = \quad \{\mathrm{AA}, \mathrm{AC}, \mathrm{AG}, \mathrm{AT}, \mathrm{CA}, \mathrm{CC}, \mathrm{CG}, \mathrm{CT}, \mathrm{GA}, \mathrm{GC}, \mathrm{GG}, \mathrm{GT}, \mathrm{TA}, \mathrm{TC}, \mathrm{TG}, \mathrm{TT}\}.$$

Let $H \in \Sigma^*_{\mathrm{DNA}}$ 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 $\mathrm{vc}(x_i, H) = \mathrm{occ}(x_i, H)$, while the *edge count* of edge $(x_i, x_j) \in E$ in H, where $x_i \sigma = \gamma x_j$, is $\mathrm{ec}((x_i, x_j), H) = \mathrm{occ}(x_i \sigma, H)$. The *order-w word count vector* $\theta^{wcv}_w(H)$ of H is the 4^w-vector

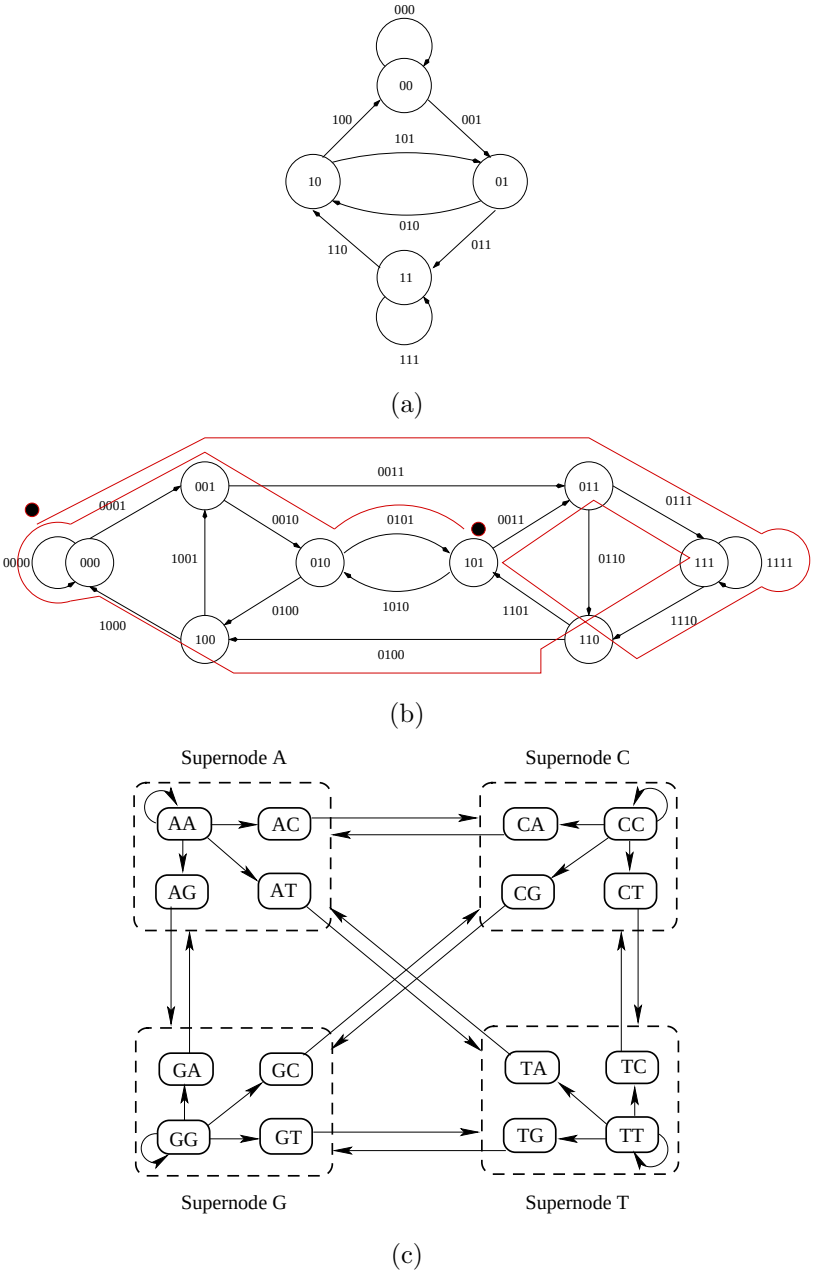


Figure 2.1: Schematic of de Bruijn graphs. (a) The order-2 de Bruijn graph over Σ_{B}. (b) The order-3 de Bruijn graph over Σ_{B}; the red line indicates a walk in the graph traced by the sequence 0001110111000101. (c) Representation of the de Bruijn graph $\mathcal{DB}^2$ over Σ_{DNA} in terms of supernodes and superedges. Each supernode consists of the 4 nodes with the same 1-symbol prefix in their labels and is closed by a dotted boundary. An edge from a node to a supernode represents a set of edges from the node to all nodes in the supernode. For example, the edge from node AC to supernode C represents the set of edges $\{(\mathrm{AC}, \mathrm{CA}), (\mathrm{AC}, \mathrm{CC}), (\mathrm{AC}, \mathrm{CG}), (\mathrm{AC}, \mathrm{CT})\}$.

Table 2.1: Nucleotide counts and frequencies in various genomic sequences.

Sequence	**Length**	**Nucleotide Counts (Frequencies)**			
		A	C	G	T
E. coli	4639675	1142228 (0.246)	1179554 (0.254)	1176923 (0.254)	1140970 (0.246)
C. pneumoniae	1229858	363689 (0.296)	249149 (0.203)	249836 (0.203)	367115 (0.299)
B. burgdorferi	910724	323079 (0.355)	130760 (0.144)	129646 (0.142)	327196 (0.359)
A. thaliana, Chr 1	30268597	9711178 (0.321)	5436538 (0.180)	5422303 (0.179)	9698578 (0.320)
S. cerevisiae, Chr 12	1078173	330586 (0.307)	207778 (0.193)	207064 (0.192)	332745 (0.309)
Uniform[1]	4000000	1001610 (0.250)	999091 (0.250)	1000543 (0.250	998756 (0.250

having components $\mathrm{occ}\,(x_i, H)$, in lexicographic order. The corresponding *order-w word frequency vector* is the 4^w-vector having components $\mathrm{freq}\,(x_i, H)$, in lexicographic order. In Figure 2.1(b), for instance, the word count vector is $\langle 2, 2, 1, 2, 1, 2, 2, 2\rangle$. Nucleotide frequencies vary between species, while, as Fickett et al. [39] observe, the frequencies of A's and T's (and hence of G's and C's) are approximately constant within a single genome. This is illustrated in Table 2.1.

Now consider the Markov chain underlying the above de Bruijn graph $\mathcal{DB}^w = (\mathcal{S}^w, E)$. The said 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)*. Here, we use DBCs in modeling of genomic signatures, based on the following intuition. Let $\mathcal{DC}$ be an order-w DBC with $4^w \times 4^w$ 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 [37]. 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

$$p_{ij} \quad = \quad \begin{cases} \frac{\mathrm{ec}((x_i,x_j),H)}{\mathrm{occ}(x_i,H)} & \text{if } \mathrm{occ}\,(x_i, H) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

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. Thus, we assume that all DBC are ergodic and hence that there is a unique *stationary distribution* $\pi = (\pi_i)$ on $\mathcal{S}^w$ satisfying $\pi P = \pi$ [37]. Ergodicity may not hold in the case of a short genomic sequence consisting of systematic repeats of a small number of length-w words.

[1]Sequence generated by a Markov Chain with uniform transition probabilities.

2.4 Genomic signatures

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 the signature. In Chapter 5, we define several signatures computed from the structure of the DBC and evaluate these and other signatures, such as the word frequency vector (θ^{wfv}) and the dinucleotides odds ratio signature (θ^{dor}) [52, 53]. In Chapter 6 we study the behavior of the θ^{dbc} signature and present associated empirical results [53].

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 $\text{ec}((x_i, x_j), H)$, where $x_i, x_j \in \mathcal{S}^w$. Let $\psi \geq 0$ be an integer *threshold*. Let $E^{\leq \psi} = \{(x_i, x_j) \in E \mid \text{ec}((x_i, x_j), H) \leq \psi\}$, 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 $\Xi = \max\{\text{ec}((x_i, x_j), H) \mid (x_i, x_j) \in E\}$ 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 4^w while the number of connected components increases from 1 to 4^w. The *vertex deletion order* θ^{vdo} is the permutation of $\mathcal{S}^w$ giving the order in which vertices become isolated during edge deletion. Let ψ_i be the smallest integer such that $\mathcal{DB}^w(H)$ has precisely i connected components. The *component-based edge deletion vector* θ^{ced} is the 4^w-vector whose i^{th} component is the number of edge deletions required to go from $i-1$ to i components. The *vertex-based edge deletion vector* θ^{ved} is the 4^w-vector whose i^{th} component is the number of edge deletions required to go from $i-1$ to i isolated vertices. The *ordered vertex-based edge deletion vector* θ^{oed} is the 4^w-vector whose i^{th} component is the total number of edge deletions required to isolate the vertex x_i, where x_i is the i^{th} element of $\mathcal{S}^w$ in lexicographic order. Define the *ordered vertex isolation frequency vector* θ^{ovif} as the 4^w-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 $2 \cdot 4^w$-vector $\pi_w \cdot \theta^{ovif}_w / 4^{w-1}$, where π_w is the estimated stationary distribution for the order-w de Bruijn chain and '$\cdot$' represents vector concatenation. Figure 2.2 illustrates the construction of the θ^{dbc} signature.

For example, consider the *E. coli K12* genome. Table 2.2 contains the order-2 transition matrix for this sequence.

For the given transition matrix, the order-2 stationary distribution is

$\langle 0.0730\ 0.0552\ 0.0511\ 0.0668\ 0.0698\ 0.0584\ 0.0747\ 0.0511\ 0.0576\ 0.0827\ 0.0584\ 0.0552\ 0.0457\ 0.0576\ 0.0698\ 0.0730 \rangle$.

The corresponding θ^{ovif}_2 signature is

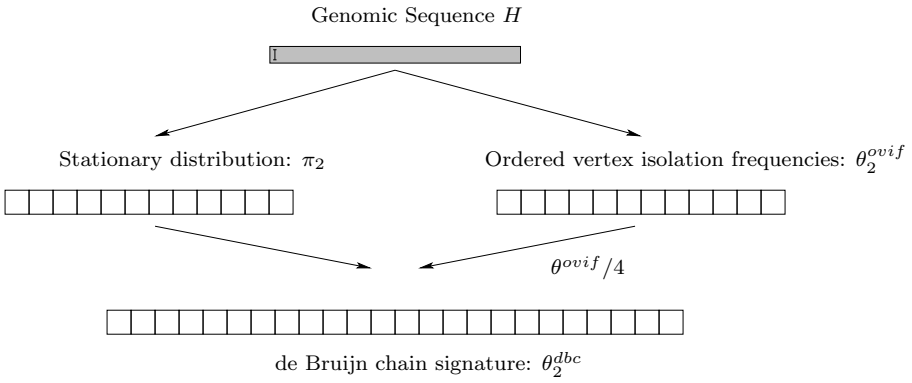


Figure 2.2: Construction of the θ^{dbc} signature.

Table 2.2: Order 2 transition matrix for the *E. coli* genome.

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0.322	0.243	0.187	0.245	0	0	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	0	0.237	0.340	0.213	0.210	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.205	0.279	0.247	0.269
0.237	0.205	0.321	0.237	0	0	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	0	0.206	0.332	0.251	0.211	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.114	0.179	0.438	0.268
0.313	0.204	0.159	0.324	0	0	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	0	0.207	0.342	0.176	0.275	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.205	0.213	0.259	0.322
0.325	0.248	0.127	0.299	0	0	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	0	0.259	0.295	0.265	0.181	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.203	0.247	0.227	0.323

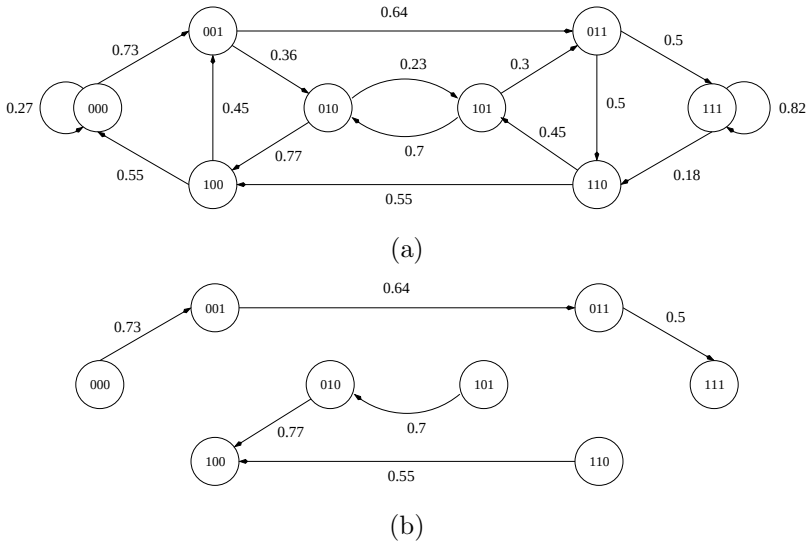


Figure 2.3: Construction of the θ^{ovif} signature from an ed ge cover. (a) The binary de Bruijn graph of order 3. (b) The edge cover from which values for individual components of θ_3^{ovif} are taken.

$$\langle 0.325\ 0.291\ 0.340\ 0.324\ 0.321\ 0.321\ 0.332\ 0.438\ 0.324\ 0.342\ 0.342\ 0.322\ 0.325\ 0.313\ 0.438\ 0.323 \rangle.$$

Therefore, the θ_2^{dbc} signature for this species is

$$\langle 0.073\ 0.0552\ 0.0511\ 0.0668\ 0.0698\ 0.0584\ 0.0747\ 0.0511\ 0.0576\ 0.0827\ 0.0584\ 0.0552\ 0.0457\ 0.0576\ 0.0698\ 0.0730$$
$$0.081\ 0.073\ 0.085\ 0.081\ 0.080\ 0.080\ 0.083\ 0.106\ 0.081\ 0.085\ 0.085\ 0.081\ 0.081\ 0.078\ 0.109\ 0.081 \rangle.$$

Our results (Chapter 6) indicate that the performance of the θ^{dbc} signature in predicting the origin of short DNA sequences is much better than the individual performances of the π and θ^{ovif} signatures or the individual performances of any of the signatures described before. The individual components of the θ_w^{ovif} signature can also be visualized 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 2.3 illustrates this point.

We compare a pair of signatures by computing the Pearson correlation coefficient [115] between them. Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from $+1$ to -1. A correlation of $+1$ means that there is a perfect positive linear relationship between variables. A correlation of -1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables. For vectors X and Y of length n, the

Pearson correlation coefficient is computed as:

$$\mathrm{R}(X,Y) = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}.$$

A pair of vectors can also be compared by calculating the distance between their transition probability matrices. The following distance measures have been used in this work. Let x, y be vectors of n elements each.

The *Bray-Curtis or Sorensen distance* between x and y is computed as

$$d_{bc}(x,y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)}. \tag{2.1}$$

If all x_i, y_i are positive, $0 \leq d_{bc}(x,y) \leq 1$. A distance of 0 indicates equal sequences.

The *Kullback-Leibler distance* is a directed discrepancy measure between two functions. It is defined as the "information" lost when a function g is used to approximate a function f, where f, g are discrete functions, and is given by

$$d_{KL}(f,g) = \sum_{i=1}^{n} f_i log\left(\frac{f_i}{g_i}\right). \tag{2.2}$$

f_i gives the true probability of the i^{th} outcome, while g_i gives the approximating probability. $d_{KL}(f,g)$ is not necessarily equal to $d_{KL}(g,f)$. Because of this asymmetry, d_{KL} is not a metric. $d_{KL}(x,y)$ is calculated similarly.

The *L1-distance or Manhattan distance* between x and y is computed as

$$d\,(x,y) = \sum_{i=1}^{n} |x_i - y_i|. \tag{2.3}$$

It takes values in the interval $[0, \infty)$.

The *L2-distance or Euclidean distance* between x and y is computed as

$$d_{L2}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{2.4}$$

It takes values in the interval $[0, \infty)$.

The *Cosine distance* between x and y is computed as

$$d_{cos}(x,y) = 1 - \frac{\sum_{i=1}^{n} (x_i y_i)}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}. \tag{2.5}$$

It takes values in the interval $[0, 1]$.

Chapter 3

Problem Definition

In this chapter, the central computational problem addressed in this dissertation is described. We define "genomic signature" as a computational concept and characterize mathematical structures that conform to this definition.

As defined in Chapters 1 and 2, a "genomic signature" is a mathematical structure that is efficiently computable from a genomic sequence at hand, sufficiently representative of the sequence from which it is derived, sufficiently different from the genomic signatures derived from genomic sequences of other genomes, and requiring significantly smaller storage space than the sequence itself.

Consider a signature function θ that takes a genomic sequence as input and returns a signature as output. Let $S_{\mathcal{G}} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ be a set of N genomes. Recall that each genome can be comprised of multiple chromosomal sequences. Consider a genome $\mathcal{G}_i \in S_{\mathcal{G}}$. Let $S_{\mathcal{G}_i} = \{H_1, \ldots, H_{N_{\mathcal{G}_i}}\}$ be the set of genomic sequences constituting genome $\mathcal{G}_i$. Consider a genomic subsequence $S \sqsubset H_q \in S_{\mathcal{G}_i}$ of length n. Then $\theta(S)$ should satisfy the following properties:

1. $\theta(S)$ *should be sufficiently representative of the sequences in* $S_{\mathcal{G}_i}$. Mathematically, this means that for a large range of n values, the distance between $\theta(S)$ and $\theta(H_r)$ for some $H_r \in S_{\mathcal{G}_i}$ should be very small.

2. $\theta(S)$ *should be sufficiently conserved within the genome* $\mathcal{G}_i$. This means that, for any genomic subsequence $S' \sqsubset H_r \in S_{\mathcal{G}_i}$ of length m, for a large range of values of m, the distance between $\theta(S)$ and $\theta(S')$ should be very small.

3. $\theta(S)$ *should be sufficiently different from the signatures of genomic sequences sampled from other*

genomes. For any genomic subsequence $S'' \sqsubset S_{\mathcal{G}_j}$, $i \neq j$, $\mathcal{G}_j \in S_{\mathcal{G}}$, the distance between $\theta(S)$ and $\theta(S'')$ should be large, much larger in magnitude than the distance between $\theta(S)$ and $\theta(S')$. It is notable here that the distance between $\theta(S)$ and $\theta(S'')$ is generally dependent on their phylogenetic separation.

4. $\theta(S)$ *should be efficiently computable*. The complexity of computing the signature will already have a large component contributed by the input sequence that is to be scanned. So, the computation process for the signature should be as inexpensive as possible.

5. $\theta(S)$ *should require much less space than the input sequence*. Since $\theta(S)$ is a predefined mathematical structure, a small, constant space requirement is ideal.

A direct application of genomic signatures is the process of identifying the origin of a DNA sequence of unknown origin. A genomic sequence S whose origin is unknown is at hand. A database $\mathcal{D}$ of a pre-defined genomic signature θ is precomputed. The origin of S is predicted by computing $\theta(S)$ and comparing it with the signatures in $\mathcal{D}$ using a pre-defined algorithm.

In the rest of this work, we introduce several genomic signatures and evaluate them with respect to origin prediction.

Chapter 4

Literature Review

In this chapter, we explore all types of genomic signatures proposed, derived, and used in the scientific literature. As discussed in Chapter 1, the term "genomic signature" has been used in two broad contexts. In the first context, the term refers to unique imprints captured from the DNA sequences of a genome $\mathcal{G}$ that have the power to distinguish between sequences sampled from $\mathcal{G}$ and sequences sampled from other genomes. In the second context, it refers to *gene expression signatures*, which are distinct conserved models of gene expression patterns observed in a set of genes during specific biological phenomena or environmental conditions [67, 84]. We will limit ourselves to the discussion of genomic signatures in the first context. The signatures we will discuss in this chapter are the dinucleotide odds ratio signature θ^{dor}, the word count vector signature θ_w^{wcv}, the word frequency vector signature θ_w^{wfv}, and Chaos Game Representation (CGR) images.

4.1 Dinucleotide odds ratio

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 [18, 22, 27, 31, 33, 38, 64, 68, 70, 106, 123, 126, 127]. Karlin and Burge [68] 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 $XY \in \Sigma^2_{\text{DNA}}$ by

$$\rho_{XY}(H) \quad = \quad \frac{\text{freq}\,(XY, H)}{\text{freq}\,(X, H)\,\text{freq}\,(Y, H)},$$

where $\text{freq}(x, H)$ is the frequency of string x as a substring in H. As an example, consider the following DNA sequence

$$S = \text{ACGATACAGATCGATACGATACACCCCAAAATTTGGGAGAGAGAGAGGGG.}$$

S has length 50. The frequencies of the mononucleotides A, C, G, and T are $\text{freq}(\text{A}, S) = 19/50 = 0.38$, $\text{freq}(\text{C}, S) = 9/50 = 0.18$, $\text{freq}(\text{G}, S) = 15/50 = 0.30$, and $\text{freq}(\text{T}, S) = 7/50 = 0.14$. The frequencies of the dinucleotides in lexicographic order are

$$\langle 0.0612, 0.1020, 0.1224, 0.1020, 0.0408, 0.0612, 0.0612, 0, 0.1837, 0, 0.1224, 0, 0.0612,$$
$$0.0204, 0.0204, 0.0408\rangle.$$

As an example, the odds-ratio corresponding to the dinucleotide AC is computed as follows:

$$\begin{aligned} \theta^{dor}_{\text{AC}}(S) &= \frac{\text{freq}(\text{AC}, S)}{\text{freq}(\text{A}, S)\,\text{freq}(\text{C}, S)} \\ &= \frac{0.1020}{(0.38)(0.18)} \\ &= 1.4912. \end{aligned}$$

So, the dinucleotides odds ratio signature $\theta^{dor}(S)$ is given by

$$\begin{aligned} \theta^{dor}(S) &= \langle \rho_{\text{AA}}(S), \rho_{\text{AC}}(S), \rho_{\text{AG}}(S), \rho_{\text{AT}}(S), \rho_{\text{CA}}(S), \rho_{\text{CC}}(S), \rho_{\text{CG}}(S), \rho_{\text{CT}}(S), \\ &\quad \rho_{\text{GA}}(S), \rho_{\text{GC}}(S), \rho_{\text{GG}}(S), \rho_{\text{GT}}(S), \rho_{\text{TA}}(S), \rho_{\text{TC}}(S), \rho_{\text{TG}}(S), \rho_{\text{TT}}(S)\rangle \\ &= \langle 0.4238, 1.4912, 1.0737, 1.9173, 0.5965, 1.8889, 1.1333, 0, 1.6114, 0, \\ &\quad 1.36, 0, 1.1504, 0.8095, 0.4857, 2.0816\rangle. \end{aligned}$$

Karlin and Burge observe that ρ values are similar throughout a genome and compare θ^{dor} for a number of organisms. They also note that the variations in the dinucleotide abundances within a genome are limited and propose its use as a genomic signature for discriminating genomic DNA. Figure 4.1 plots the θ^{dor} signatures for 5 species.

Karlin et al. [70] 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. Given the dinucleotides odds ratio signatures $\theta^{dor}(S_1)$ and $\theta^{dor}(S_2)$ for two sequences S_1 and S_2, respectively, the δ-distance between them is computed as follows:

$$\delta(\theta^{dor}(S_1), \theta^{dor}(S_2)) = \frac{1000}{16} d\left(\theta^{dor}_{S_1}, \theta^{dor}_{S_2}\right).$$

Karlin et al. [70] compare and contrast genome-wide compositional biases and distributions of short oligonucleotides across 15 diverse prokaryotes that have substantial genomic sequence collections. They observe that

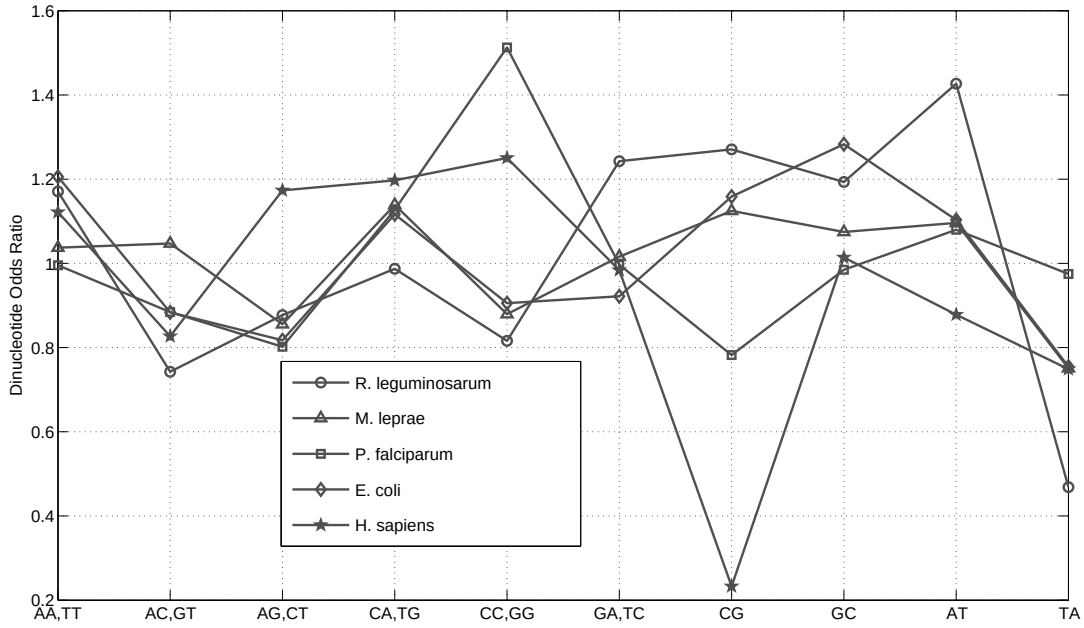


Figure 4.1: Plot of the θ^{dor} signatures for 5 species. The species shown are *R. leguminosarum*, *M. leprae*, *P. falciparum*, *E. coli*, and *H. sapiens* chromosome 21, using double stranded genomic DNA sequences. Note the symmetry in the odds-ratios of dinucleotides that are reverse complements of one another.

the dinucleotide relative abundance profiles over multiple 50 Kb disjoint contigs within the same genome are approximately constant. They further note that the differences between θ^{dor} vectors of 50 Kb sample contigs of different genomes almost always exceed the differences between those of the same genomes.

Campbell et al. [16] compare θ^{dor} signatures of prokaryotic, plasmid, and mitochondrial DNA. Their comparisons of θ^{dor} signatures for plasmids, both specialized and broad-range, and their hosts indicate that plasmids and their hosts have substantially compatible (similar) genome signatures. They also observe that while mammalian mitochondrial (Mt) genomes are very similar, and animal and fungal Mt are generally moderately similar, they diverge significantly from plant and protist Mt sets. [1] They find that in terms of similarities between θ^{dor} signatures, archaea are not a coherent clade and contain some greatly divergent species. They also found no consistent pattern of signature differences among thermophiles. They group prokaryotes by environmental criteria (e.g., habitat propensities, osmolarity tolerance, chemical conditions) and do not observe any tight correlations of genome signatures in these groups. This does not provide evidence for the proposition of Karlin et al. [70] that dinucleotide composition could be related to the determination of behaviors of species to various environmental conditions.

Gentles and Karlin [42] 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*. They find that dinucleotide relative abundances are remarkably constant across human chromosomes and within the DNA of a particular species. They also observe that "dinucleotide biases differ between species, providing a genome signature that is characteristic of the bulk properties of an organism's DNA".

Jernigan and Baran [64] analyze 22 sequences, representing 19 species, to assess stability of the signature in windows ranging in size from 50 kilobases down to 125 bases. For each sequence, they compute the distance of the global signature from the locally-computed signatures for all non-overlapping windows on each sequence. They find that these distances are log-normally distributed with nearly constant variance and with means that tend to zero slower than reciprocal square root of window size. Further, the mean distance within genomes is larger for protist, plant, and human chromosomes, and smaller for archaea, bacteria, and yeast, for any window size. They 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.

[1] Protists are eukaryotic organisms that vary a lot from one another and are classified into the kingdom Protista.

4.2 Chaos Game Representations (CGRs) of sequences

The second most widely-used method in the literature to visualize and study the composition of DNA sequences is Chaos Game Representation (CGR). Mathematically, the Chaos Game is an iterated function system. CGR uses a two-dimensional heat-map-style plot to provide a visual representation of composition of a given DNA sequence in terms of DNA word frequencies. The tiled geometrical patterns in the CGR sharpen with increasing DNA word lengths. Visualization of DNA sequence composition using CGRs was first proposed by Jeffrey [63]. Subsequent work on CGRs involved mathematical characterization of CGRs to predict the presence or absence of a sequence in any gene family by using properties of CGRs as classifiers of gene families [35], analysis of CGR images to deduce that CGRs within a species were closely similar while the differences between CGRs of two sequences grew with increased phylogenetic separation [58], and calculation of entropic profiles of DNA sequences through analysis of their CGRs [94]. In 1993, Goldman [44] asserted that dinucleotide and trinucleotide frequencies alone explain the patterns observed in CGRs. In 2004, Wang et al. [129] challenged Goldmann's results. They concluded that "if a CGRs resolution is $1/2^k$ and the DNA sequence is much longer than k, this CGR is completely determined by all the numbers of length k oligonucleotide occurrences". They also mention that the θ^{dor} signature and CGRs are related and that all genomic signatures are members of a spectrum of properties where each signature has its own properties. In this work, distances between CGR images are also suggested as a basis for phylogeny.

Deschavanne et al. [31] have devised a graphic method which makes it possible to show frequencies of the various words of a given length, by a CGR image with a fractal structure. Each pixel of the image is dedicated to a word and the pixel intensity is proportional to the frequency of its associated word in the genome. The resolution of the image determines the length of the studied words. At the lowest level are the frequencies of mononucleotides A, C, G, and T. The image showing mononucleotide frequencies has the smallest resolution; four pixels with each pixel allocated to one mononucleotide. This is illustrated by the first image in Figure 1 in Deschavanne et al. [31], with the top-left square allocated to the frequency of C, and the top-right, bottom-right, and bottom-left squares allocated to the frequencies of G, T, and A, respectively. For every higher resolution-level, each pixel is broken up into four pixels while prefixing a character from Σ_{DNA} to it. The four corners are assigned in the exact same order of the prefixed characters as described above for the mononucleotide scenario. For example, sequences of 2 bases generated starting from the pixel for C are CC, GC, AC, and TC starting at the top-left square and moving clockwise one square at a time. Thus a 128 pixels-square image will describe frequencies of all the 7 letter-words in the analyzed genomic sequence. The CGR images for oligonucleotides from length 1 to length 8 are presented in the same figure in Deschavanne et al. [31]. The dictionary of all the words of fixed length used in a genome can therefore be visualized by a CGR image.

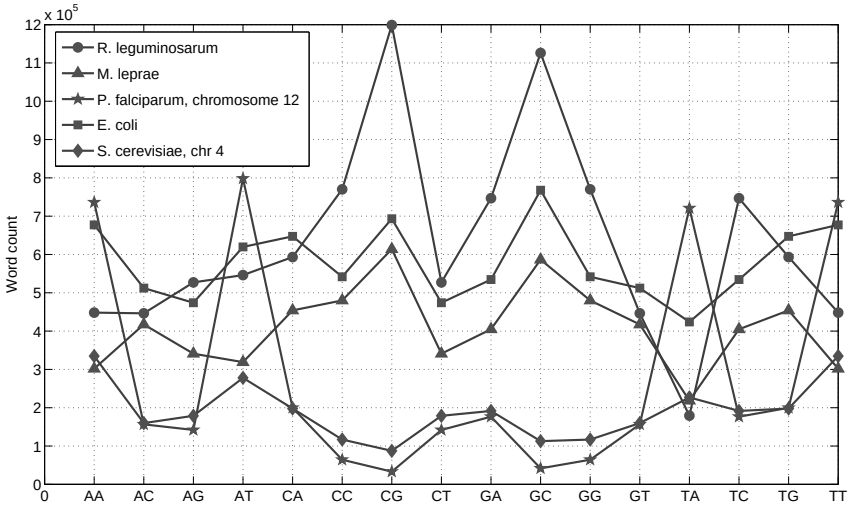


Figure 4.2: Word count θ_2^{wcv} signatures for five diverse species.

Deschavanne et al. [31, 38] have constructed CGR images from oligonucleotide frequencies and built the application GENSTYLE, 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.

In the first figure at the GENSTYLE website, Deschavanne et al. [29] use CGR images for 7-long DNA words to illustrate that the CGR image may vary significantly from one species to another. While some images are well-structured, others are chaotic with no obvious well-defined structure. Also some regions of the image can be astonishingly denser than the remaining regions. In the fourth figure, Deschavanne et al. [29] discuss the diversity and conservation of the CGR image as a signature. They compare the CGR image for length 5 DNA words for the entire genome of a species to CGR images of fragments of different lengths sampled from that genome. They observe that as the fragment size decreases, so does the resemblance of its CGR image to the CGR image for its entire genome.

4.3 Word count vector signatures θ^{wcv}

The simple word count vector has also been used as a genomic signature in various works. Figure 4.2 illustrates the θ_2^{wcv} signatures of 5 diverse species.

For bacterial species, Coenye and Vandamme [22] correlate the δ-distance (Section 4.1) with 16S rDNA sequence similarity and DNA-DNA hybridization values. They demonstrate that the correlation between the genomic signature and DNA-DNA hybridization values is high, while the overall correlation between the genomic signature and 16S rDNA sequence similarity is low, except for closely related organisms (16S rDNA similarity $> 94\%$). For 57 prokaryotic genomes, Sandberg et al. [106] quantify the species-specificity of θ_2^{wfv} genomic signatures in the complete genomes of 57 prokaryotes. They confirm that the θ_2^{wfv} genomic signature is genome-wide, with high species-specificity in both coding and non-coding regions, and compare G+C content, oligonucleotide frequency, and codon bias. Dufraigne et al. [33] and van Passel et al. [126] employ oligonucleotide frequencies to identify regions of horizontal gene transfer (HGT) in prokaryotes. Carbone et al. [18] correlate the ecological niches of 80 Eubacteria and 16 Archaea to codon bias used as a genomic signature.

The application TETRA [123] uses tetranucleotide frequencies to calculate similarity between sequences. It stores tetranucleotide usage patterns in all genomic sequences available at NCBI. Based on a Markov model, it evaluates the levels of over- and underrepresentation for each of the 256 possible tetranucleotides in a submitted DNA sequence. These data are then normalized via a z-transformation and their correlation coefficients to tetranucleotide frequencies of existing genomes are calculated. TETRA is available both as a stand-alone tool and a web-based system at `http://www.megx.net/tetra_new/index.html`.

4.4 Gene fragments as genomic barcodes

DNA barcoding is a method of characterizing an organism using a relatively short subsequence of its genome at an agreed upon position. Prof. Paul D. N. Hebert was the first to propose the use of a specific genomic region as a genomic barcode. Hebert et al. [55] have established that a 648 base long region in subunit I of the mitochondrial gene cytochrome c oxidase I (COI) can serve as the core of a global bioidentification system for higher animals. This is based on the concept that most eukaryotic cells contain mitochondria, and hence, mitochondrial DNA (mtDNA), which has a fast mutation rate resulting in greater variation of mtDNA between species and much smaller variation of mtDNA within a species.

At a level of finer detail the term DNA barcode differs slightly from the classical definition of a barcode. All pieces of a product being sold or displayed are marked by the same 11-digit barcode identifier. However, the same cannot be said about the DNA barcode of different individuals from the same species. Although a DNA barcode is meant to be species-specific, there is some variation, albeit small, of DNA barcodes within members of a species.

Hebert et al. [57] tested the effectiveness of the 648-bp region of the mitochondrial gene, cytochrome c oxidase I (COI) in discriminating bird species. They determined COI barcodes for 260 North American birds and found that each species has a different COI barcode. Also, the differences between closely related species were, on average, 18 times higher than the differences within species. Their results identified 4 probable new species of American birds, and they suggested that a global survey with a standard screening of sequence difference of COI sequences can identify new organisms quickly and inexpensively.

Next, Hebert et al. [56] applied DNA barcodes to the common neotropical skipper butterfly *Astraptes fulgerator*. They used a combination of natural history and morphological studies along with DNA barcoding of museum specimens to show that *A. fulgerator* is a complex of at least 10 different cryptic butterflies species [2] from northwestern Costa Rica. However, Brower [15] challenged the results of Hebert et al. [56] by conducting experiments that showed that at least 3, but not more than 7 mtDNA clades that may correspond to the above cryptic species are supported by the evidence. He also addressed some methodological and philosophical weaknesses of Hebert's claim that the DNA barcoding approach could serve as a proxy for the arduous, painstaking work of genuine systematics. Thus, it can be concluded that the process of delimiting species using DNA barcoding is much dependent on the premises and analytical methods used by the researchers. Hence, the results vary and are subjective.

Subsequently, Smith et al. [114] examined whether the cytochrome COI DNA barcode could function as a tool for species identification and discovery for the 20 morphospecies of *Belvosia* parasitoid flies (Diptera: Tachinidae) that have been reared from caterpillars (Lepidoptera) in Area de Conservacio' n Guanacaste (ACG), northwestern Costa Rica. They also found that barcoding not only discriminates among all 17 highly host-specific morphospecies of ACG Belvosia, but it also raises the species count to 32 by revealing that each of the three generalist species are actually arrays of highly host-specific cryptic species. In 2007, Smith et al. [113] DNA-barcoded 2134 flies belonging to what appeared to be the 16 most generalist of the reared tachinid morphospecies. They encountered 73 mitochondrial lineages separated by an average of 4% sequence divergence and, as these lineages are supported by collateral ecological information, and, where tested, by independent nuclear markers (28S and ITS1), the authors therefore viewed these lineages as provisional species. Each of the 16 initially apparent generalist species were categorized into one of four patterns: (i) a single generalist species, (ii) a pair of morphologically cryptic generalist species, (iii) a complex of specialist species plus a generalist, or (iv) a complex of specialists with no remaining generalist. In sum, there remained 9 generalist species classified among the 73 mitochondrial lineages analyzed.

In [108], marine biologists Schander and Willassen propose that DNA barcodes could be a very useful tool

[2] *Cryptic species* are animals that appear morphologically identical but are genetically quite distinct.

for taxonomy. They argue that barcodes could help to identify cryptic and polymorphic species and give means to associate life history stages of unknown identity and provide tools for higher taxonomic resolution of disparate life forms in case of ambiguous morphology. However, they conclude that morphology and other biological information about species is vital for their identification and cannot be made obsolete by barcodes.

In plants, the same COI gene could not be used as a barcode because of a much slower rate of COI gene evolution in higher plants than in animals. In 2005, Kress et al. [71] proposed the nuclear internal transcribed spacer region and the plastid trnH-psbA intergenic spacer as potentially usable DNA regions for applying barcoding to flowering plants. They based their proposition on the fact that internal transcribed spacer is the most commonly sequenced locus used in plant phylogenetic investigations at the species level and shows high levels of interspecific divergence. The trnH-psbA spacer, although short ($\approx$ 450-bp), is the most variable plastid region in angiosperms and is easily amplified across a broad range of land plants. They compared the total plastid genomes of tobacco and deadly nightshade enhanced with trials on widely divergent angiosperm taxa, including closely related species in seven plant families and a group of species sampled from a local flora encompassing 50 plant families (for a total of 99 species, 80 genera, and 53 families), and found results that suggest that the sequences in this pair of loci have the potential to discriminate among the largest number of plant species for barcoding purposes. In 2008, however, Savloleinen and his team [72] undertook intensive field collections in two biodiversity hotspots (Mesoamerica and southern Africa). Using >1600 samples, they compared eight potential genomic regions for ideal DNA barcoding properties. They assessed to what extent a "DNA barcoding gap" is present between intra- and interspecific variations, using multiple accessions per species. They identified a portion of the plastid matK gene that had an adequate rate of variation, easy amplification, and alignment, as a good DNA barcode for flowering plants. They also analyzed >1000 species of Mesoamerican orchids, and DNA barcoding with matK alone revealed cryptic species and proved useful in identifying species listed in the Convention on International Trade of Endangered Species (CITES) appendixes.

The prospect of cataloging ancient life using DNA barcodes was explored by Lambert et al. [73] in 2005. They sequenced the 5' terminus of the mitochondrial COI gene of individuals belonging to the moa of New Zealand, a group of extinct ratite birds. They derived precise information about the number of moa species that existed using a phylogenetic approach based on a large data set including protein coding and 12S DNA sequences, as well as morphology. They showed that each moa species had a distinct COI barcode and that the variation in COI barcodes among individuals of any moa species was low. They suggested that DNA barcoding might also help detect other extinct animal species and build a large-scale directory of ancient life.

Taxonomists saw the application of DNA barcodes to classify species as an oversimplification of rigorous

systematic taxonomic methods. It has been found that using an mtDNA barcode to assign a species name to an animal will be ambiguous or erroneous some 23% of the time [86]. Studies with insects suggest an equal or even greater error rate, due to the frequent lack of correlation between the mitochondrial genome and the nuclear genome or the lack of a barcoding gap [132]. Given that insects represent over 75% of all known organisms [56], this suggests that while mtDNA barcoding may work for vertebrates, it may not be effective for the majority of known organisms.

The Consortium for Barcoding of Life (CBOL) has built the Barcoding of Life Database (BOLD) [102]. BOLD provides a repository for barcode records coupled with analytical tools and serves as an online workbench for the DNA barcode community. It also provides a species identification tool that accepts DNA sequences from the barcode region and returns a taxonomic assignment to the species level when possible.

4.5 Classification of DNA fragments by different methods

The genomic signatures described in this chapter have been demonstrated by investigators to vary very little within a species and much more between species. While between-species variations in signatures have been amply pointed out in the literature, systematic tests to determine the power of these signatures in predicting the origin of short DNA fragments have been rare. In this section, we summarize all efforts made so far in scientific literature towards exploring the power of genomic signatures described in this chapter in the classification of DNA fragments of unknown origin.

A few tests were conducted by Deschavanne and his team [31, 30] to study the conservation of the CGR genomic signature in DNA fragments. They observe that "images obtained from parts of a genome present the same structure as that of the whole genome" (Fourth figure in [29]). They note that, with decreased segment sizes, a distinct reduction in sharpness of the images is observed. However, their overall structure resembles the general design of the image for the genome. They carried out an analysis using 13 species: *A. fulgidus*, *B. subtilis*, *C. acetobutylicum*, *D. radiodurans*, *E. coli*, *H. sapiens*, *M. leprae*, *M. musculus*, *M. tuberculosis*, *S. cerevisiae*, *S. pombe*, *T. maritiana*, and *V. cholerae*. They randomly selected a 100 kb subsequence from each species. The euclidean distances between the signature of the 100 kb subsequence of one species and the signatures of the 100 kb subsequences of other species were computed. The mean distance of each species to other species is indicated by the stars in Figure 3(b) in Deschavanne et al. [31]. The quantile plots and the associated mean and standard deviations represent the distribution of distances between signatures of intra-genomic fragments. Deschavanne et al. [31] observe that the mean intra-genomic distance is about 700, compared with 1900 for the mean inter-genomic distance. The right-most quantile plot in the figure represents the distribution of distances between signatures of inter-genomic fragments.

Deschavanne et al. [30] examine the effect of DNA word length and fragment length on the classification efficiency using CGR images. The results are summarized in Figure 1 in [30]. The authors used 16 unspecified genomes for the experiment. They randomly sampled series of (100, 25, 10, 5, 1, 0.5 and 0.1 kb) non-overlapping DNA fragments from each genome and generated the corresponding CGR images. They observe that whole genome patterns are maintained even when the segment size decreases. In order to test to what extent short DNA fragments share properties of the species they derive from, the authors use an unsupervised clustering technique to group fragments as a function of the characteristics of their word distribution. The proportion of well-classified fragments is then computed by comparing the fragments to the origin. They observe that the variability of word frequencies within a genome decreases when the size of the fragment increases. They also note that longer words may be more species-specific although their frequency along the genome may be more variable. The proportion of well-classified fragments roughly increases with the size of the fragment, whatever the length of the words considered in the calculation. They, however, point out that a surprising high proportion of 1 kb fragments is properly classified. Similarly, the proportion of well-classified fragments increases with the length of the words, whatever the size of the fragment. Their key observation is that, while a perfect classification is already achieved using 3 letter-words and 100 kb fragments, the analysis of longer words is required to get good results when small fragments are considered. They comment that the usage of long words (5-letters) seems to be quite constant along each genome while being also very species-specific.

Next, we explore the literature that has utilized dinucleotide compositions to classify DNA fragments. Nakashima et al. [89] analyzed human, yeast, and *E. coli* coding sequences in terms of dinucleotide occurrences. In 16-dimensional space, they observed that the human and *E. coli* clusters were distinctly separated while that for yeast was positioned in between. They observed that genes from the same organism were clustered in the space. Later, Nakashima et al. [90] used the 16 normalized dinucleotide compositions to analyze the protein-encoding nucleotide sequences in nine complete genomes including 3 Gram-negative bacteria (*Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*), 2 Gram-positive bacteria (*Mycoplasma genitalium*, *Mycoplasma pneumoniae*), one cyanobacterium (*Synechocystis sp.*), two archaea (*Methanococcus jannaschii*, *Archaeoglobus fulgidus*), and one eukaryote (*Saccharomyces cerevisiae*). They extracted protein-coding nucleotide sequences from these species using sequences and feature tables from the DNA data bank of Japan. They observed that, as expected, the dinucleotide composition is significantly different between the organisms. They found that using the dinucleotide composition alone the genes from the nine complete genomes cluster around their respective genomes' centers with 80% accuracy in dinucleotide composition space using Euclidean distances to compute separation. They observe that the fact that the whole genome compositions are close to compositions of coding regions suggests that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. Despite all of the above

observations, this work did not report experimental results on the effectiveness of the genomic signature in identifying targets accurately.

Karlin et al. [70] established differences among θ^{dor} signatures of 15 diverse prokaryotes. They observe the constancy of θ^{dor} signatures of 50 kb sequences sampled from any genome. They also observe that the separations between θ^{dor} signatures of sequences sampled within a genome were much less on average than the separations between θ^{dor} signatures of sequences sampled from different genomes. Even though Karlin et al. [70] point out interesting oligonucleotides of varying lengths in different species, they do not comment on the efficacy of the θ^{dor} signature on target identification or the effects of available sequence length and word length on it. Campbell et al. [16] point out the similarities in dinucleotide compositions of plasmids and their hosts but do not address the problem of target identification in their work, or the effect of the lengths of sequences available from hosts and plasmids on the similarity between them. Gentles and Karlin [42] make the same observations, while examining the genomes of yeast, arabidopsis, and drosophila, but do not mention anything formally about the mapping of short sequences to their respective origins. Karlin et al. [69] show that δ-distances between different genomic sequences in the same species are low and are generally smaller than the between-species δ-distances. They point out extremes in short oligonucleotide over- and under-representation in several species. They assess homogeneity of the dinucleotide relative abundance profile through the delta-distance and propose the following standards for measuring the similarity of an available sequence to a putative target:

$0 < \delta < 15$	random
$15 < \delta < 30$	very close
$30 < \delta < 45$	close
$45 < \delta < 65$	moderately related
$65 < \delta < 95$	distantly related

Based on this standard, Jernigan and Baran [64] examine 22 sequences from 19 species and 17 genera with the understanding that the sequence is fundamentally non-stationary, exhibiting statistically significant variations in base frequencies between non-overlapping 50 kb windows in a genome. They state that the scaled δ-distance ($\delta/\sqrt{n}$), where n is the length of the sequence, is a statistical invariant for any benchmark sequence generated by a Markov chain exhibiting the same signature as the given sequence. Their results suggest that profiles seen through smaller windows are statistically closer to the global signature. The profiles seen through larger windows tend toward the signature but local fluctuations tend to zero slower than $1/\sqrt{n}$ (i.e., the convergence rate is "sub-Markov"). For details, please see Figure 1 in [64].

All of the above results further reinforce the conserved nature of dinucleotide odds within a genome, but do

not say much about inter-genomic distances and how they contribute to origin prediction.

4.6 Summary

The signatures described in this chapter demonstrate that signatures differ among species, but with the exception of the CGR images, none formally address the amount of variation, identification of unknown DNA, and the effect of short available sequence length on these signatures. Moreover, they do not examine sequences from the point of view of the structure of a graph on which the sequence can be defined as a walk. 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 [96, 97, 134, 101, 135]. 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 (See Chapter 5). We introduce the de Bruijn chain signature θ^{dbc} and demonstrate that it performs better than all existing genomic signatures with emphasis on target identification from short DNA segments (See Chapter 6). This signature performs much better than oligonucleotide frequency vectors in differentiating among diverse genomes. 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.

Chapter 5

Purely graph-based genomic signatures

5.1 Introduction

A genomic sequence H of length n defines a walk over a de Bruijn graph $\mathcal{DB}^w$ of order w. The de Bruijn graph $\mathcal{DB}^w$ over the DNA alphabet $\Sigma_{\text{DNA}} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$ has vertex set Σ_{DNA}^w and edge set $\Sigma_{\text{DNA}}^{w+1}$. Using a sliding window that moves one character at a time, vertex counts and edge counts of $\mathcal{DB}^w(H)$ can be recorded using occurrences of w-mers in H. In this chapter, we explore various properties of the graph $\mathcal{DB}^w(H)$ and structures that are conserved across sequences sampled randomly from a genome and vary between sequences sampled randomly from different genomes.

5.2 Databases of organisms

To test the accuracy of these first, purely graph-based signatures in identifying an organism from its sequence, we used a database of diverse genomic sequences of various lengths, including α-proteobacteria, infectious bacteria, and eukaryotes. Table 5.1 identifies these genomic sequences and the acronyms used for them in this chapter.

Table 5.1: List of genomic sequences in the set of diverse species. Bacterial and eukaryotic species have been used.

Species	Acronym	Sequence length	NCBI identifier
Rhizobium leguminosarum	RL	5.1 Mb	NC_008380
Erythrobacter litoralis	EL	3.1 Mb	NC_007722
Mycobacterium leprae	ML	3.3 Mb	NC_002677
Neisseria meningitidis	NM	2.2 Mb	NC_008767
Plasmodium falciparum	PF	chr 12, 2.3 Mb	NC_004316
Pseudomonas aeruginosa	PA	6.4 Mb	NC_002516
Streptococcus pneumoniae	SP	2.1 Mb	NC_008533
Escherichia coli	EC	4.7 Mb	NC_000913
Caenorhabditis elegans	CE	chr 1, 15.3 Mb	NC_003279
Homo sapiens	HS	chr 1, 228.7 Mb	AC_000044
Arabidopsis thaliana	AT	chr 4, 18.85 Mb	NC_003075
Saccharomyces cerevisiae	SC	chr 4, 1.6 Mb	NC_001136

5.3 The word count (frequency) vector signature θ^{wcv} (θ^{wfv})

The word count vector signature θ^{wcv} of order w is defined as the 4^w long vector whose i^{th} component is given by $\text{occ}\,(x_i, H)$, where x_i is the i^{th} word in lexicographic order in Σ^w_{DNA}. The word frequency vector signature θ^{wfv} of order w is defined as the 4^w long vector whose i^{th} component is given by $\text{occ}\,(x_i, H)/(n-w+1)$, where x_i is the i^{th} word in lexicographic order in Σ^w_{DNA}. Consider the DNA sequence S = AAACGAGTCATTCCTGAGGAGCACC. Here $n = 25$. The corresponding $\theta_2^{wcv}(S)$ signature is

$$\langle 2, 2, 3, 1, 2, 2, 1, 1, 3, 1, 1, 1, 0, 2, 1, 1\rangle,$$

and since $n - w + 1 = 24$, the corresponding $\theta_2^{wfv}(S)$ signature is

$$\langle 0.0833, 0.0833, 0.125, 0.0417, 0.0833, 0.0833, 0.0417, 0.0417, 0.125, 0.0417, 0.0417, 0.0417, 0, 0.0833, 0.0417, 0.0417\rangle.$$

In Chapter 4, the word count and word frequency vector and its variants have been discussed in detail, so we will not discuss its fundamental properties further. In the rest of this section, we propose a mathematical framework within which we characterize the separation between the θ^{wfv} signatures of sequences generated by the same DBC. We also characterize the separation between the θ^{wfv} signatures of sequences generated by different DBCs. Thereafter, we describe empirical results that illustrate the conservation of the θ^{wfv} signature within a genome. We also study the accuracy of the θ^{wfv} signature in origin prediction.

5.3.1 Mathematical results for θ^{wfv}

Let $\mathcal{DC}$ be an ergodic, order-w DBC. Let H be a sequence generated by $\mathcal{DC}$, where $|H| = n$. If $x_i, x_j \in \mathcal{S}^w$, the probability of transition from state x_i to state x_j is given by $p_{i,j}$, and the stationary probability for x_i is π_i.

Let $x = \sigma_1\sigma_2 \ldots \sigma_w \in \Sigma^w_{\text{DNA}}$. A *period* of x is an integer i, where $1 \leq i \leq w$, such that $x[1 \ldots i] = x[w-i+1 \ldots w]$. Two occurrences $H[i \ldots i+w-1]$ and $H[j \ldots j+w-1]$ of x in H *overlap* if $i \leq j \leq i+w-1$ or $j \leq i \leq j+w-1$. An *x-clump* in H is a maximal subsequence of one or more consecutive overlapping occurrences of x. For example, 2 is a period of $x = \text{AACAA}$, and AACAACAACAACAA is a clump with 4 occurrences of x. Waterman [130] notes that the count of a rare DNA word in H is a function of the number of x-clumps in H, which approximately follows a Poisson distribution [130], with parameter λ_β (derived below). Let x be a DNA word with shortest period d. Then a *declumping* event with respect to x is defined as the event of not observing the string $x' = x[1 \ldots d]$. Suppose the probability of occurrence of x' is p_x. Then the probability of a declumping event is given by $q_x = 1 - p_x$. The number of occurrences of x within a clump is approximately geometric with mean $1/p_x$ [130].

Lemma 5.1. *Let X_x be the random variable that is the number of occurrences of word x in genomic sequence H. Then the probability generating function of X_x is*

$$f_{X_x}(t) = \exp\left(\frac{\lambda_x(t-1)(1-p_x)}{1-q_x t}\right).$$

Proof. Let Z be the random variable that is the number of x-clumps in H, and let C_i be the number of occurrences of x in the i^{th} clump. Hence,

$$X_x = \sum_{i=1}^{Z} C_i.$$

Since Z has (approximately) a Poisson distribution with parameter λ_x, the probability generating function for Z is

$$f_Z(t) = \sum_{k=0}^{\infty} e^{-\lambda_x} \frac{(\lambda_x t)^k}{k!} = e^{\lambda_x(t-1)}.$$

The probability generating function for each C_i is

$$f_C(t) = p_x \sum_{k=0}^{\infty} (q_x t)^k = \frac{p_x}{1-q_x t}.$$

Assuming independence of the C_i, the probability generating function for X_x is

$$f_{X_x}(t) = f_Z(f_C(t)) = \exp\left(\lambda_x\left(\frac{p_x}{1-q_x t} - 1\right)\right) = \exp\left(\frac{\lambda_x(t-1)(1-p_x)}{1-q_x t}\right).$$

□

Lemma 5.2. $\mathbf{E}[X_x] = \dfrac{\lambda_x q_x}{p_x}$ *and* $\mathrm{Var}[X_x] = \dfrac{\lambda_x q_x}{p_x}\left(\dfrac{2q_x}{p_x}+1\right)$.

Proof. By results in [37], $\mathbf{E}[X_x] = f'_{X_x}(1)$ and $\mathrm{Var}[X_x] = f''_{X_x}(1) + f'_{X_x}(1) - (f'_{X_x}(1))^2$.

$$f'_{X_x}(t) \;=\; \exp\left(\frac{\lambda_x(t-1)(1-p_x)}{1-q_x t}\right)\left(\frac{\lambda_x(1-p_x)}{1-q_x t} + \frac{q_x\lambda_x(t-1)(1-p_x)}{(1-q_x t)^2}\right).$$

$$\begin{aligned} & f''_{X_x}(t) \\ &= f'_{X_x}(t)\left(\frac{\lambda_x(1-p_x)}{1-q_x t} + \frac{q_x\lambda_x(t-1)(1-p_x)}{(1-q_x t)^2}\right) + f_{X_x}(t)\left(\frac{2q_x\lambda_x(1-p_x)}{(1-q_x t)^2} + \frac{2q_x^2\lambda_x(t-1)(1-p_x)}{(1-q_x t)^3}\right). \end{aligned}$$

We have

$$\mathbf{E}[X_x] = f'_{X_x}(1) = \frac{\lambda_x q_x}{p_x}.$$

Now,

$$f''_{X_x}(1) \;=\; \frac{\lambda_x^2 q_x^2}{p_x^2} + \frac{2\lambda_x q_x^2}{p_x^2}.$$

Therefore,

$$\begin{aligned} \mathrm{Var}[X_x] &= f''_{X_x}(1) + f'_{X_x}(1) - (f'_{X_x}(1))^2 \\ &= \frac{\lambda_x^2 q_x^2}{p_x^2} + \frac{2\lambda_x q_x^2}{p_x^2} + \frac{\lambda_x q_x}{p_x} - \frac{\lambda_x^2 q_x^2}{p_x^2} \\ &= \frac{\lambda q_x}{p_x}\left(\frac{2q_x}{p_x}+1\right). \end{aligned}$$

The lemma follows from the above calculations. □

Lemma 5.3. *Let H be a genomic sequence of length n, and let χ_H^w be its word count vector. Fix threshold $\tau > 0$. Then*

$$\Pr[d(\chi, \mathbf{E}[\chi]) \geq 4^w\tau] \leq \sum_{x\in\mathcal{S}^w} \frac{n\pi_x}{\tau^2}\left(\frac{2q_x}{p_x}+1\right).$$

Proof. Let $\chi_H^w = (X_1, X_2, \ldots X_{4^w})$. Since $\mathbf{E}[X_x] = n\pi_x = (\lambda_x q_x)/p_x$, we have $\lambda_x = (n\pi_x p_x)/q_x$. The distance between χ and $\mathbf{E}[\chi]$ is $d(\chi, \mathbf{E}[\chi]) = \sum_{x\in\mathcal{S}^w} |X_x - \mathbf{E}[X_x]|$. By Chebyshev's bound and Lemma 5.2, we obtain

$$\Pr[|X_x - \mathbf{E}[X_x]| \geq \tau] \leq \frac{\mathrm{Var}[X_x]}{\tau^2} = \frac{\lambda_x q_x}{p_x\tau^2}\left(\frac{2q_x}{p_x}+1\right) = \frac{n\pi_x}{\tau^2}\left(\frac{2q_x}{p_x}+1\right).$$

The lemma follows from the resulting inequality:

$$\Pr[d(\chi, \mathbf{E}[\chi]) \geq 4^w\tau] \leq \sum_{x\in\mathcal{S}^w} \Pr[|x - \mathbf{E}[x]| \geq \tau].$$

□

Theorems 5.4 and 5.5 address the ability of word count vectors to identify and distinguish DBCs.

Theorem 5.4. *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 χ_1 and χ_2 be their respective order-w word count vectors. Then,*

$$\Pr\left[d\left(\chi_1, \chi_2\right)) \geq 2 \cdot 4^w \tau\sqrt{n}\right] \quad \leq \quad \frac{2}{\tau^2}(2 \cdot 4^w - 1).$$

Proof. The component-wise expected values in χ_1 and χ_2 are the same. Their expected difference is therefore the 0 vector. Therefore,

$$d\left(\chi_1 - \mathbf{E}\left[\chi_1\right], \chi_2 - \mathbf{E}\left[\chi_2\right]\right) = d\left(\chi_1, \chi_2\right).$$

Furthermore using $T = \tau\sqrt{n}$ we obtain,

$$\Pr\left[d\left(\chi_1, \mathbf{E}\left[\chi_1\right]\right) \geq 4^w T\right] \quad = \quad \Pr\left[d\left(\chi_2, \mathbf{E}\left[\chi_2\right]\right) \geq 4^w T\right].$$

Using the above equations and Lemma 5.3, we obtain

$$\Pr\left[d\left(\chi_1 - \mathbf{E}\left[\chi_1\right], \chi_2 - \mathbf{E}\left[\chi_2\right]\right) \geq 2 \cdot 4^w T\right] \quad = \quad \Pr\left[d\left(\chi_1, \chi_2\right) \geq 2 \cdot 4^w T\right].$$

$$\begin{aligned}\Pr\left[d\left(\chi_1, \chi_2\right) \geq 2 \cdot 4^w T\right] &\leq \Pr\left[d\left(\chi_1, \mathbf{E}\left[\chi_1\right]\right) \geq 4^w T\right] + \Pr\left[d\left(\chi_2, \mathbf{E}\left[\chi_2\right]\right) \geq 4^w T\right] \\ &= 2 \sum_{x \in \mathcal{S}^w} \frac{n\pi_x}{T^2}\left(\frac{2q_x}{p_x} + 1\right).\end{aligned}$$

If $x' = x[1 \ldots d]$, where d is the smallest period of x, $|x| \geq |x'|$. Therefore, $p_x \geq \pi_x$ and $\frac{q_x}{p_x} \leq \frac{1 - \pi_x}{\pi_x}$, which yields

$$\begin{aligned}\Pr\left[d\left(\chi_1, \chi_2\right) \geq 2 \cdot 4^w T\right] &\leq \frac{2}{\tau^2} \sum_{x \in \mathcal{S}^w} \pi_x \left(\frac{1 - \pi_x}{\pi_x} + 1\right) \\ &= \frac{2}{\tau^2} \sum_{x \in \mathcal{S}^w} \left(2 - \pi_x\right).\end{aligned}$$

From the above results we have

$$\Pr\left[d\left(\chi_1, \chi_2\right) \geq 2 \cdot 4^w \tau\sqrt{n}\right] \quad \leq \quad \frac{2}{\tau^2}(2 \cdot 4^w - 1).$$

□

Let H_1 and H_2 be genomic sequences of length n, generated independently by DBCs $\mathcal{DC}_1$ and $\mathcal{DC}_2$ of orders s_1 and s_2, respectively. Let $\chi_1 = \chi_1^{H_1}$ and $\chi_1 = \chi_2^{H_2}$ be their order-w word count vectors. This assumption formalizes the separation of genomic sequences obtained from different organisms.

Assumption 5.1. *There exists a non-negative real number $\gamma \in (0, 1]$ such that*

$$\Pr\left[d\left(\mathbf{E}\left[\chi_1\right], \mathbf{E}\left[\chi_2\right]\right) \geq 3 \cdot 4^w \tau\sqrt{n}\right] \geq \gamma.$$

Then, the distance $d(\chi_1, \chi_2)$ can distinguish $\mathcal{DC}_1$ and $\mathcal{DC}_2$.

Theorem 5.5. *let $X_{x,1}$ and $X_{x,2}$ denote the counts of x in H_1 and H_2, respectively. Assuming that H_1 and H_2 are both generated by Markov chains $\mathcal{DC}_1'$ and $\mathcal{DC}_2'$ of order w, let $\pi_{x,1}$ and $\pi_{x,2}$ denote the stationary probabilities of state x in $\mathcal{DC}_1'$ and $\mathcal{DC}_2'$, respectively. If there exists a constant γ as in Assumption 5.1 then,*

$$\Pr\left[d\left(\chi_1, \chi_2\right) \geq 4^w \tau\sqrt{n}\right] \quad \geq \quad \gamma - \frac{2}{\tau^2}(2 \cdot 4^w - 1).$$

Proof. Treating $d(\chi_1, \chi_2)$, $d(\chi_1, \mathbf{E}[\chi_1])$, $d(\chi_2, \mathbf{E}[\chi_2])$, and $d(\mathbf{E}[\chi_1], \mathbf{E}[\chi_2])$ as distances d, d_1, d_2, and d_3, respectively, in 1-dimensional space and using $T = \tau\sqrt{n}$ we obtain,

$$\begin{aligned} d_3 &\leq d + d_1 + d_2 \\ \Pr[d_3 \geq 3 \cdot 4^w T] &\leq \Pr[d \geq 4^w T] + \Pr[d_1 \geq 4^w T] + \Pr[d_2 \geq 4^w T]. \end{aligned}$$

From Assumption 5.1, Lemma 5.3, and $\pi_x \leq p_x$ we obtain,

$$\begin{aligned} \gamma &\leq \Pr[d(\chi_1, \chi_2) \geq 4^w T] + \sum_{x \in \mathcal{S}^w} \frac{n\pi_{x,1}}{T^2}\left(\frac{2q_{x,1}}{p_{x,1}} + 1\right) + \sum_{x \in \mathcal{S}^w} \frac{n\pi_{x,2}}{T^2}\left(\frac{2q_{x,2}}{p_{x,2}} + 1\right), \\ \Pr\left[d(\chi_1, \chi_2) \geq 4^w \tau\sqrt{n}\right] &\geq \gamma - \frac{1}{\tau^2}\sum_{x \in \mathcal{S}^w}(2 - \pi_{x,1}) - \frac{1}{\tau^2}\sum_{x \in \mathcal{S}^w}(2 - \pi_{x,2}) \\ &= \gamma - \frac{1}{\tau^2}(2 \cdot 4^w - 1) - \frac{1}{\tau^2}(2 \cdot 4^w - 1) \\ &= \gamma - \frac{2}{\tau^2}(2 \cdot 4^w - 1). \end{aligned}$$

The theorem follows. □

By Theorem 5.5, the probability that the distance between the word count vectors of sequences generated by different DBCs exceeds $4^w \tau\sqrt{n}$, increases with τ. Sequences assumed to be generated by two different DBC with sufficiently different stationary distributions would have a high probability of being separated by a large distance.

5.3.2 Empirical results for the θ^{wcv} signature

In this section, we examine the properties of the θ^{wcv} signature within genomic sequences of a diverse set of genomes.

First, we computed θ^{wcv} signatures of orders 2 and 3 for entire chromosomal sequences of AT chromosomes I, II, and III, CE chromosomes I, III, and IV, and SC chromosomes IV, V, and VIII. We then computed Pearson correlation between the signatures of each pair of chromosomes. Figure 5.1 illustrates the results. Figure 5.1(a) contains the results corresponding to order-2 θ^{wcv} signatures while Figure 5.1(b) contains the results corresponding to order-2 θ^{wcv} signatures. Each rectangle with rounded corners illustrates the Pearson correlation coefficients between the θ^{wcv} signatures of chromosomal sequences within genome. Edges between rectangles indicate the range of Pearson correlation coefficients between the θ^{wcv} signatures of each of the 9 pairs of chromosomes for every pair of organisms.

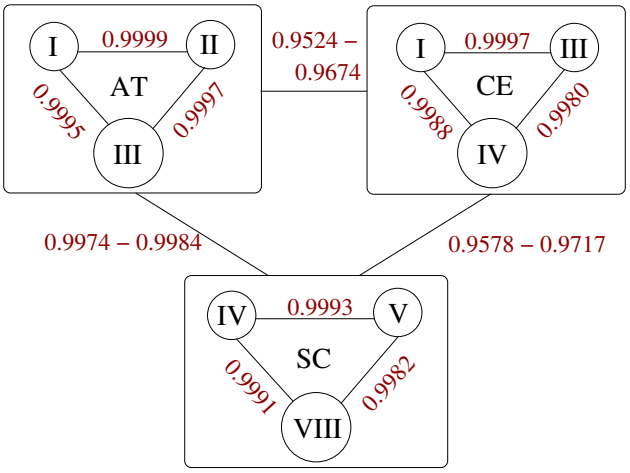
Observe that the range of Pearson correlation coefficients between the chromosomal sequences of a genome is not very different from the range of Pearson correlation coefficients between the chromosomal sequences of two different genomes irrespective of the order of the θ^{wcv} signature. It is expected of a good genomic signature that the range of Pearson correlation coefficients between genomes be lower in magnitude than the range of Pearson correlation coefficients within a genome. This property is not demonstrated by the θ^{wcv} signature as is illustrated in Figure 5.1. Therefore, we study graph-based signatures to examine if they are competitive with word-count based signatures.

Next, we examined the accuracy of the θ^{wcv} signature in predicting the origin of relatively short DNA sequences. Figure 5.2 illustrates the results. From each species on the x-axis, 100 sequence samples of of length 10 kb each were randomly sampled. For each sample, the θ_2^{wcv} signature was computed and its distance computed from each θ_2^{wcv} signature in the set of 12 genomic signatures for all species on the x-axis. The species with the closest distance was predicted as the origin of the sample. The number of correct predictions is the *accuracy*, which is plotted on the y-axis. Although the θ_2^{wcv} signature is not able to effectively distinguish between entire chromosomal sequences of different species, it can predict the origin of short DNA sequences with higher accuracy than was expected from the results in Figure 5.1.

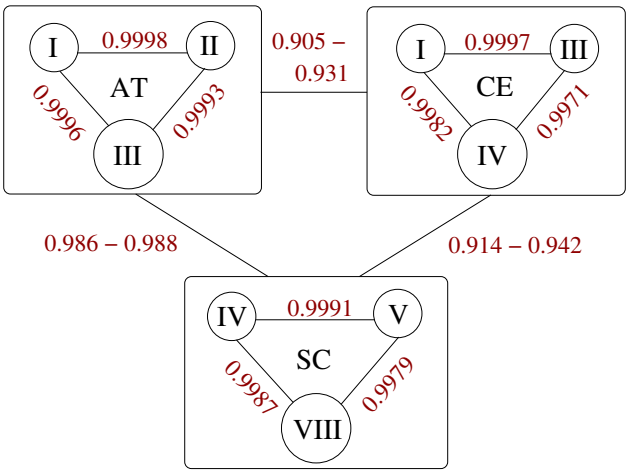
5.4 The edge deletion cycle

Let $\psi \geq 0$ be an integer *threshold*. Let $E^{\leq\psi} = \{(i,j) \in E \mid \mathrm{ec}\,((i,j),H) \leq \psi\}$ 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 $\Xi = \max\{\mathrm{ec}\,((i,j),H) \mid (i,j) \in E\}$ and deleting edges with tied counts in arbitrary order. The *ψ-edge deletion of $\mathcal{DB}^w$* is $\mathcal{DB}^w(\psi) = (\mathcal{S}\,, E - E^{\leq\psi})$. As ψ increases from 0 to Ξ, the number of connected components in $\mathcal{DB}^w(\psi)$ increases from 1 to 4^w, while the number of isolated vertices increases from 0 to 4^w.

Figures 5.4, 5.5, 5.6 and 5.7 illustrate the various stages of edge deletion in the order-3 DBC over the binary



(a)



(b)

Figure 5.1: Pearson correlation coefficients between θ^{wcv} signatures of AT, CE, and SC. θ^{wcv} signatures of orders (a) 2 and (b) 3 have been used.

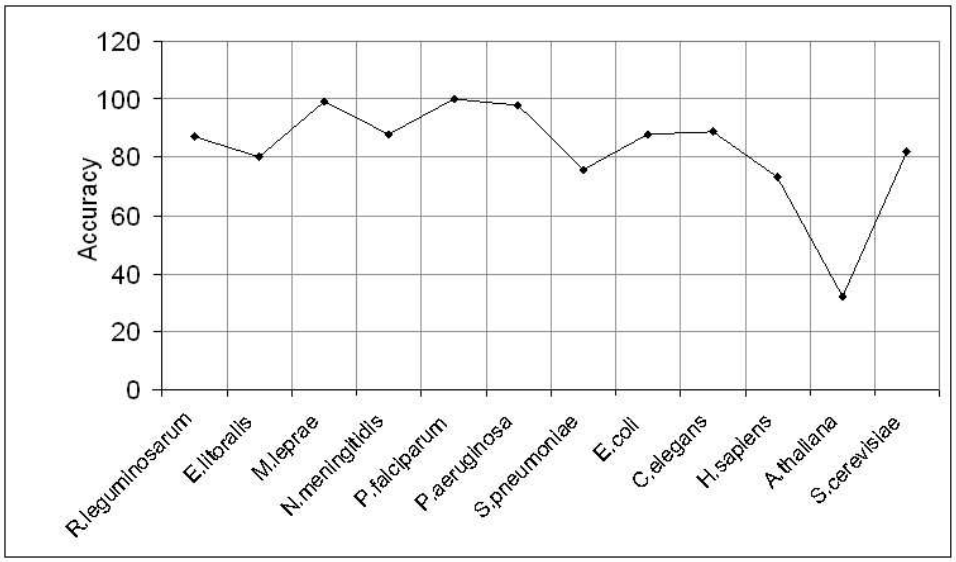


Figure 5.2: Accuracy of the θ_2^{wcv} signature. DNA fragments of length 10 kb and a database of 12 diverse species have been used.

alphabet shown in Figure 5.3.

5.5 The vertex deletion order signature θ^{vdo}

In the course of an edge deletion cycle, vertices of the DBC become isolated. The number of isolated vertices increases from 0 at the beginning of the edge deletion cycle to $|V|$ at the end of the edge deletion cycle, where $|V|$ is the number of vertices in the DBC. The *vertex deletion order* θ^{vdo} is the permutation of $\mathcal{S}^w$ giving the order in which vertices become isolated during edge deletion.

Figure 5.8 contains the graphical representations of the θ_2^{vdo} signatures of several species, including multiple chromosomes of some species. Observe that the θ_2^{vdo} signatures of different organisms are very different from each other. The θ_2^{vdo} signatures of two closely related species *C. pneumoniae* and *C. muridarum* are very similar. The θ_2^{vdo} signatures of two different AT chromosomes are almost the same, as are the θ_2^{vdo} signatures of three different SC chromosomes. Figures 5.9 and 5.10 contain the graphical representations of θ_3^{vdo} signatures of AT and SC chromosomes. Figure 5.11 (a) illustrates the Pearson correlation coefficients between the θ_2^{vdo} signatures of the 16 SC chromosomes while Figure 5.11 (b) illustrates the Pearson correlation coefficients between the θ_3^{vdo} signatures of the 16 SC chromosomes. Similarly, Figure 5.12 (a) illustrates the Pearson correlation coefficients between the θ_3^{vdo} signatures of the 5 AT chromosomes while Figure 5.12 (b)

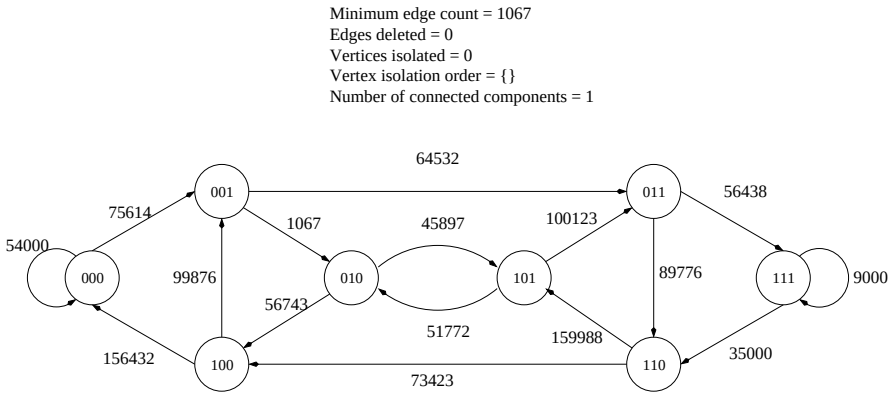


Figure 5.3: A binary DBC of order 2 with edge counts.

illustrates the Pearson correlation coefficients between the θ_3^{vdo} signatures of the 6 CE chromosomes. Observe that the θ^{vdo} signature is conserved very well between the genomic sequences of a given genome. However, while the θ^{vdo} signature is conserved very well within a genome, it is not very effective at distinguishing between genomes as is illustrated in Figure 5.13.

5.6 The component-based edge deletion vector θ^{ced}

In the course of an edge-deletion cycle, the number of connected components in the DBC increases from 1 at the beginning of the edge deletion cycle to $|V|$ at the end of the edge deletion cycle. Let ψ_i be the smallest integer such that $\mathcal{DB}^w(\psi_i)$ has precisely i connected components. The *component-based edge deletion vector* θ^{ced} is the 4^w-vector whose i^{th} component is the number of edge deletions required to go from $i-1$ to i components. Figure 5.14 shows the θ_2^{ced} signatures for the entire chromosomal sequences of several sequences. The ability of the θ_3^{ced} signature to distinguish between the chromosomal sequences of the three species AT, CE, and SC is better than that of the θ_3^{wcv} and θ_3^{vdo} signatures as is illustrated in Figure 5.15. However, the accuracy of the θ_2^{ced} signature in identifying the origin of a short DNA sequence of length 10 kb is much less than the accuracy of the order-2 signatures discussed so far as illustrated in Figure 5.16.

5.7 The ordered vertex-based edge deletion vector θ^{oed}

The *vertex-based edge deletion vector* θ^{ved} is the 4^w-vector whose i^{th} component is the number of edge deletions required to go from $i-1$ to i isolated vertices. The *ordered vertex-based edge deletion vector* θ^{oed}

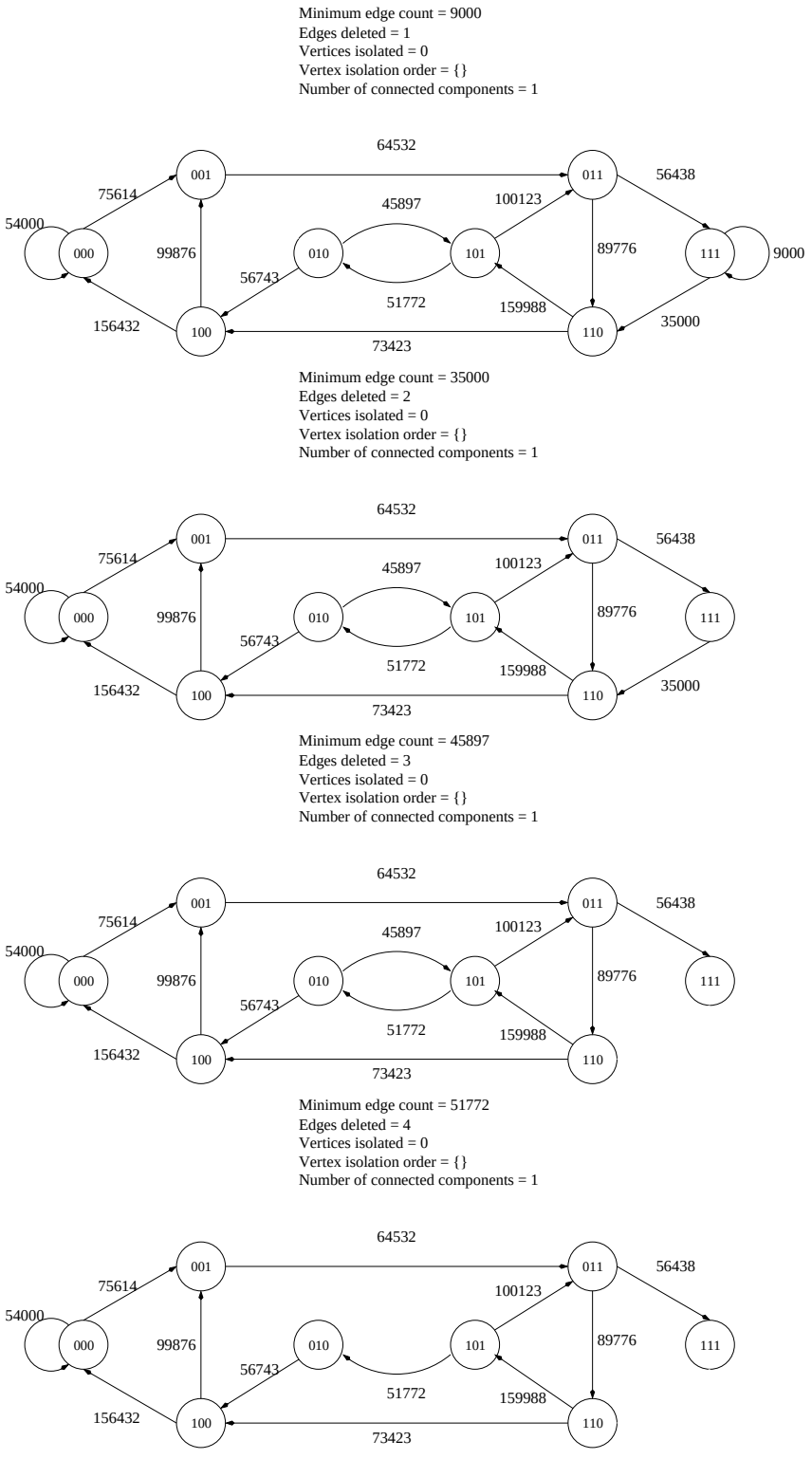


Figure 5.4: Edge deletion cycle - I. Edge deletion cycle steps: 1, 2, 3, and 4.

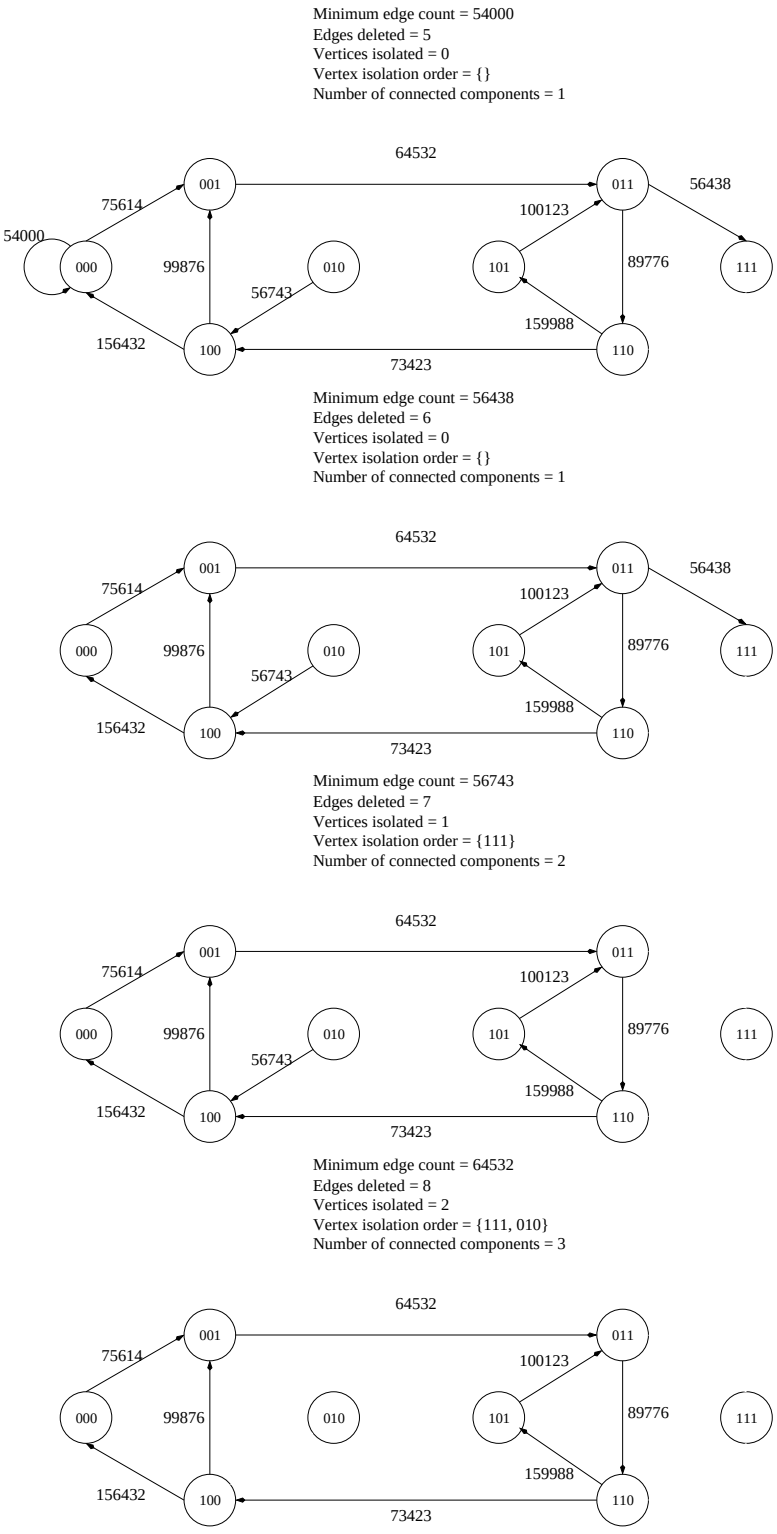


Figure 5.5: Edge deletion cycle - II. Edge deletion cycle steps: 5, 6, 7, and 8.

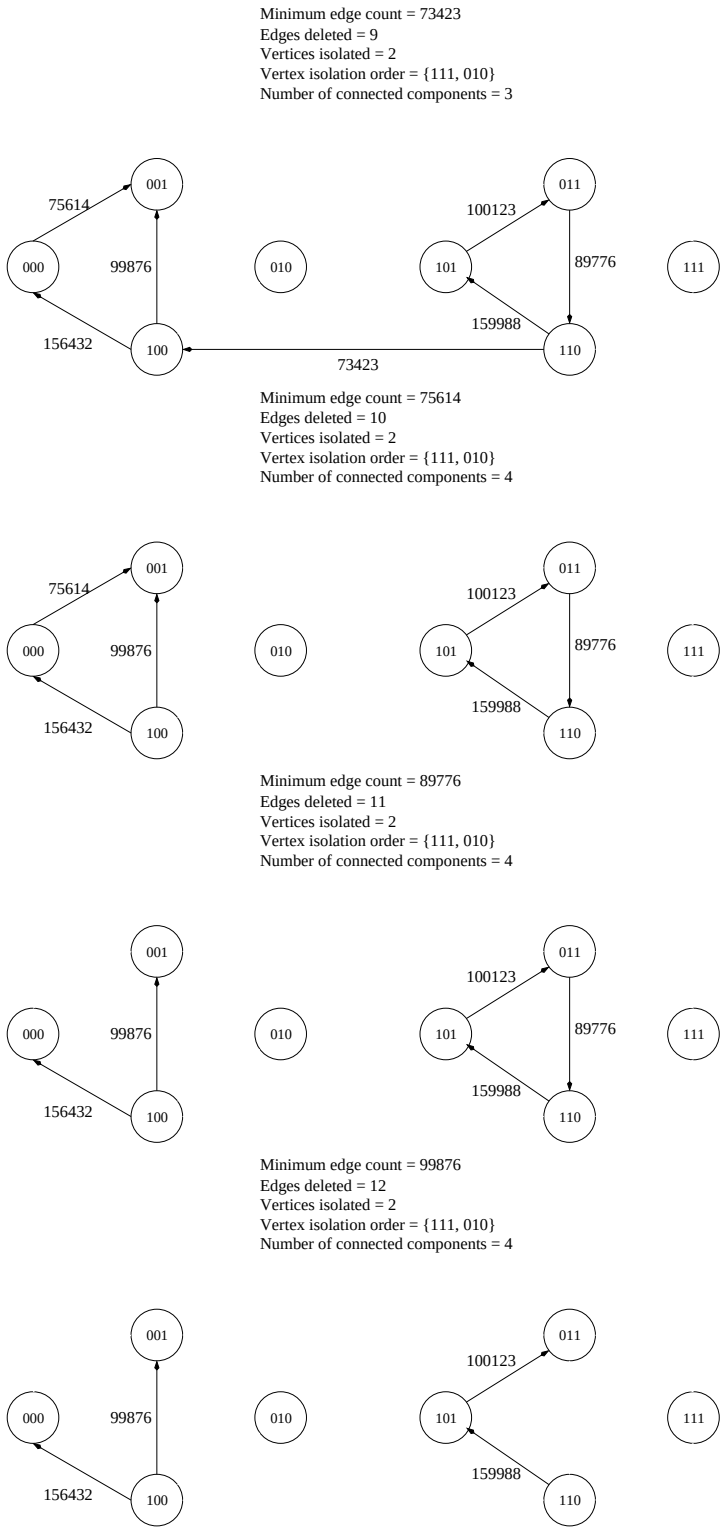


Figure 5.6: Edge deletion cycle - III. Edge deletion cycle steps: 9, 10, 11, and 12.

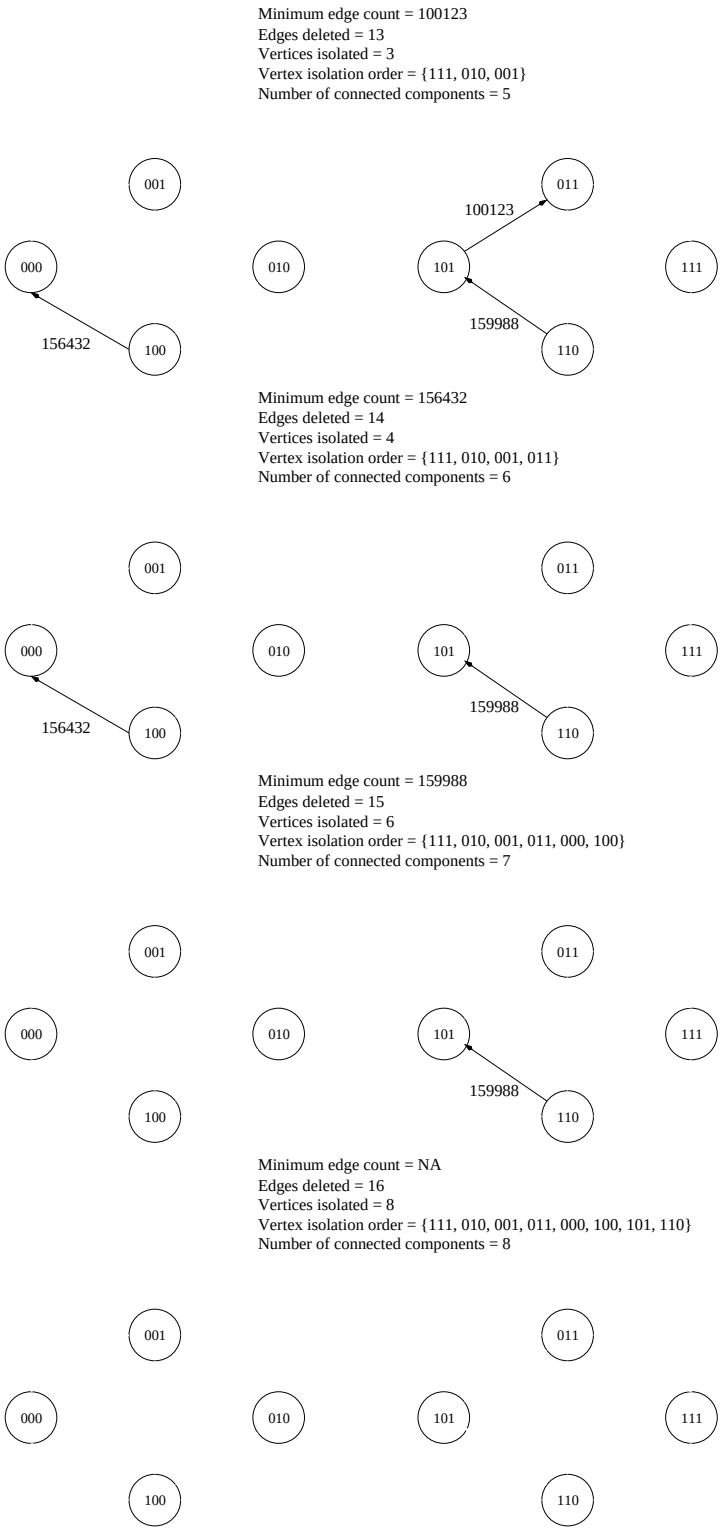


Figure 5.7: Edge deletion cycle - IV. Edge deletion cycle steps: 13, 14, 15, and 16.

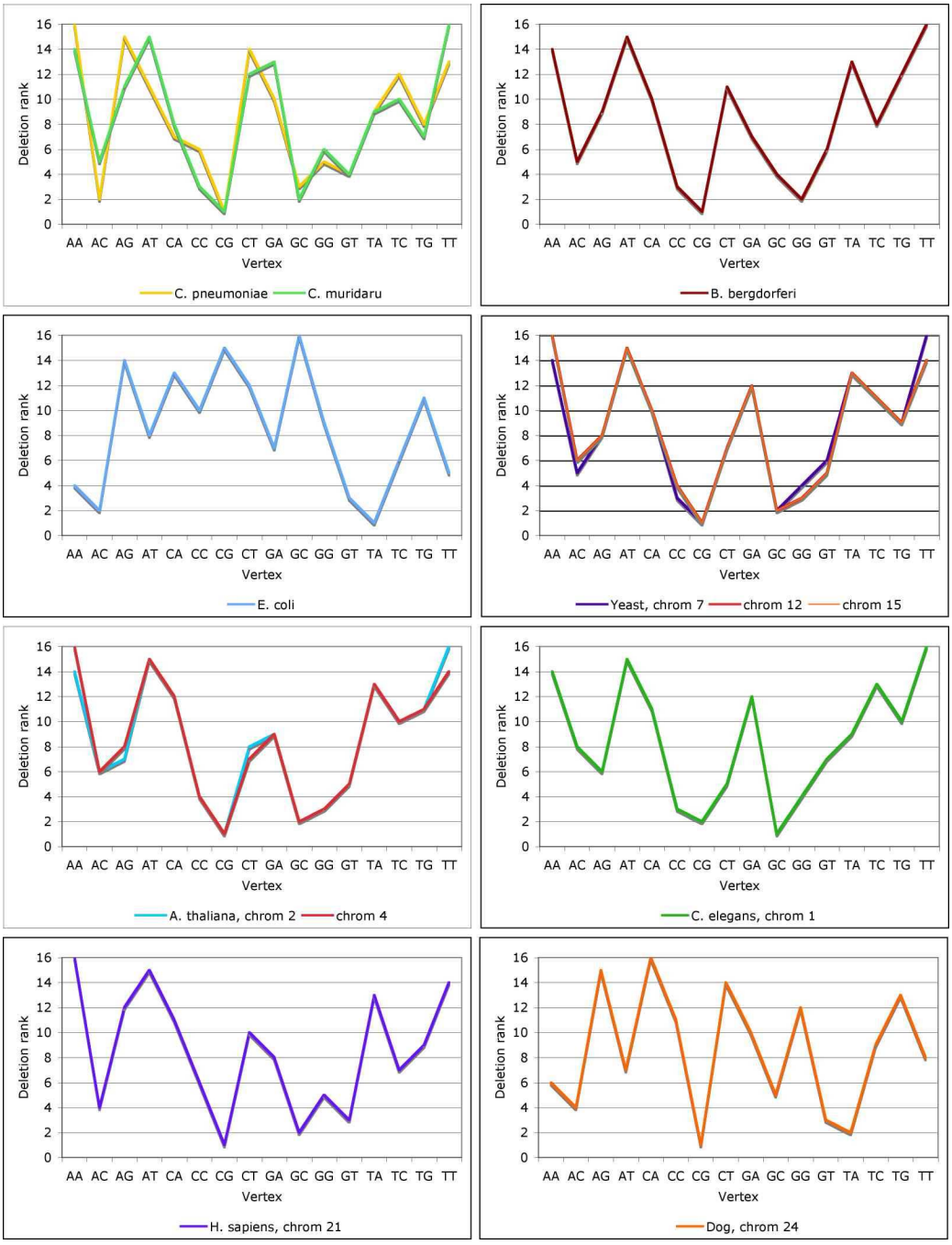


Figure 5.8: θ_3^{vdo} signatures of entire chromosomes of various species.

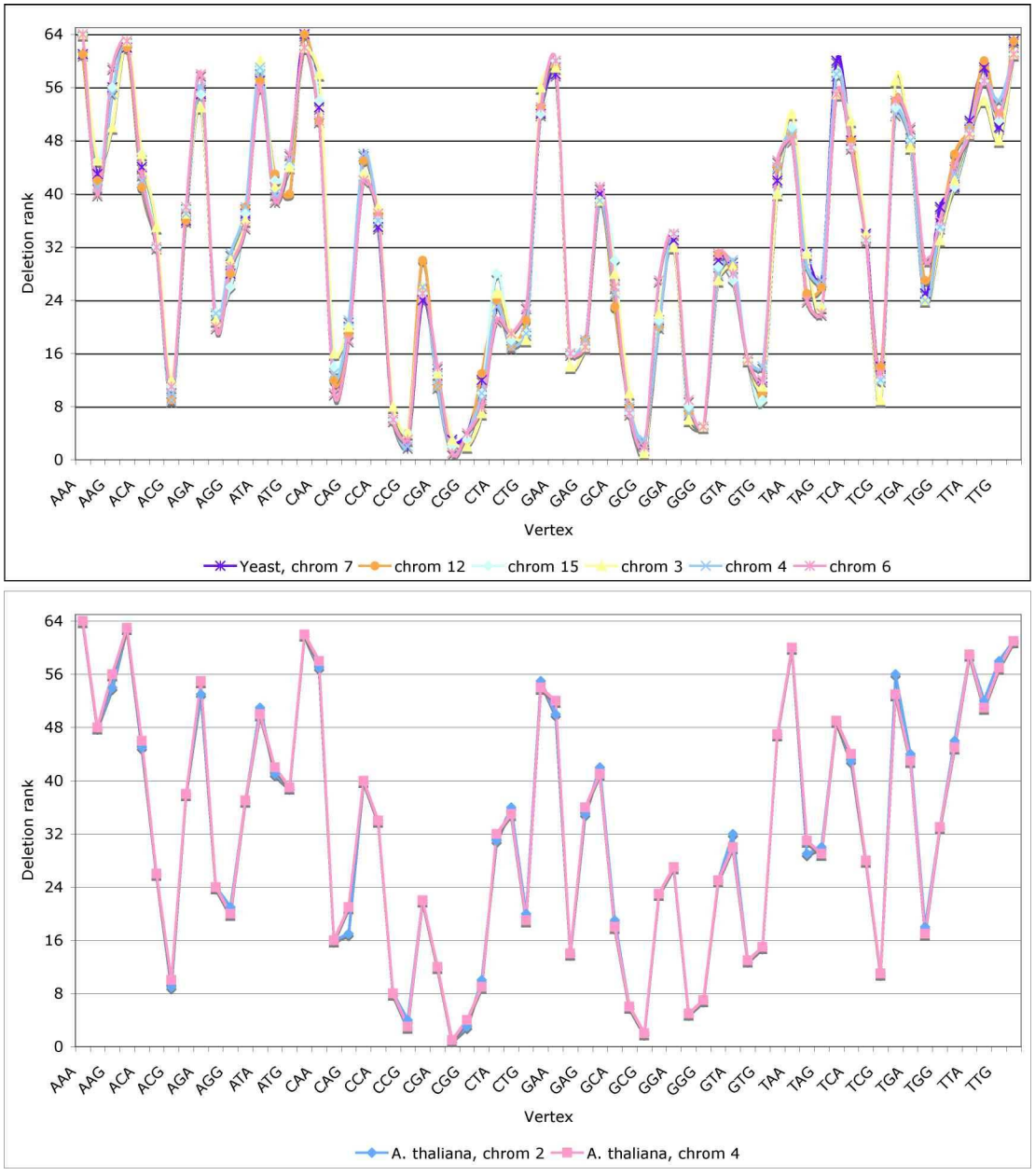


Figure 5.9: θ_3^{vdo} signatures of entire chromosomes of SC and AT.

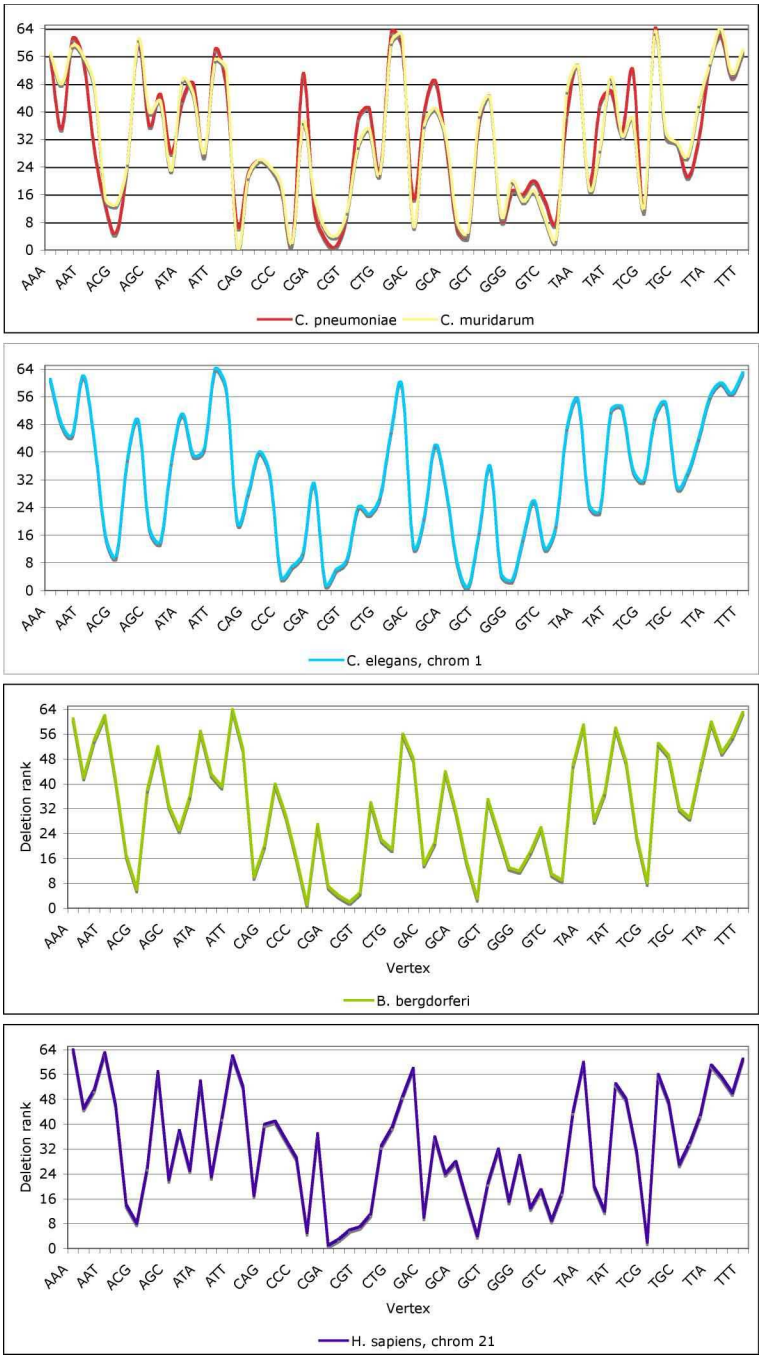
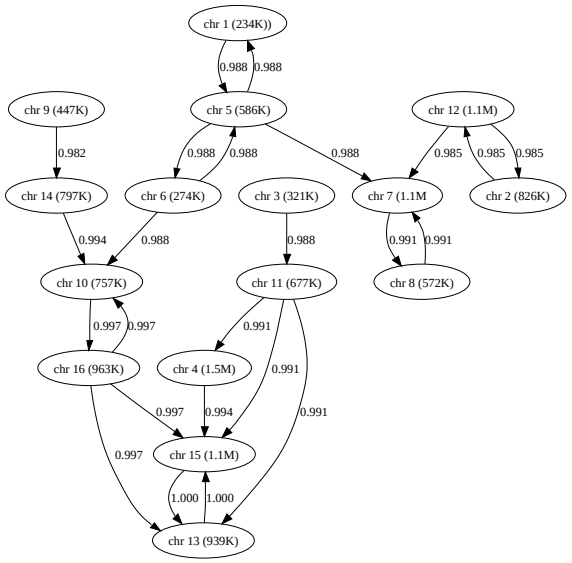
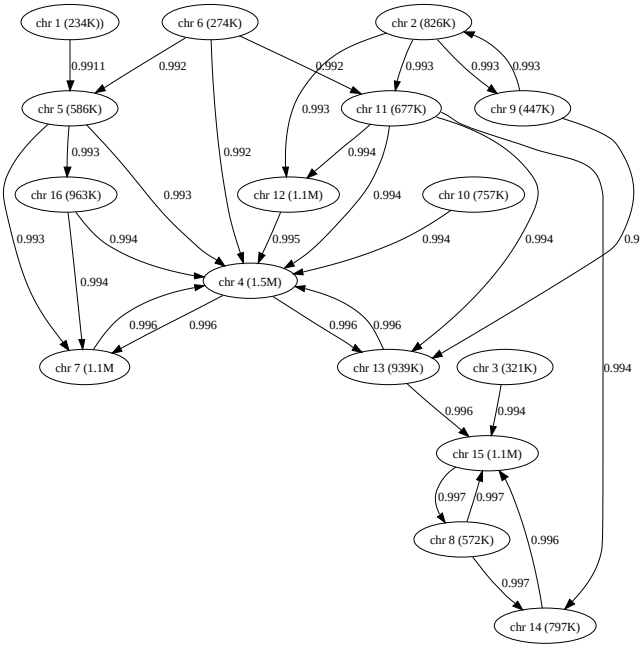


Figure 5.10: θ_3^{vdo} signatures of entire chromosomes of CP, CM, CE, BB, and HS.

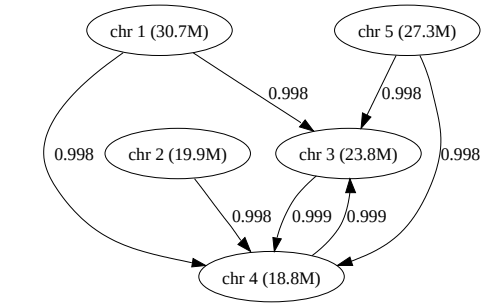


(a)

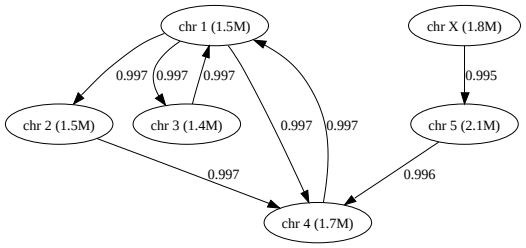


(b)

Figure 5.11: Pearson correlation coefficients between θ^{vdo} signatures of the 16 SC chromosomes. (a) θ_2^{vdo} signatures and (b) θ_3^{vdo} signatures have been used.



(a)



(b)

Figure 5.12: Pearson correlations between θ_3^{vdo} signatures of chromosomes of (a) AT and (b) CE. Chromosomal sequences of the 5 AT chromosomes and 6 CE chromosomes have been used.

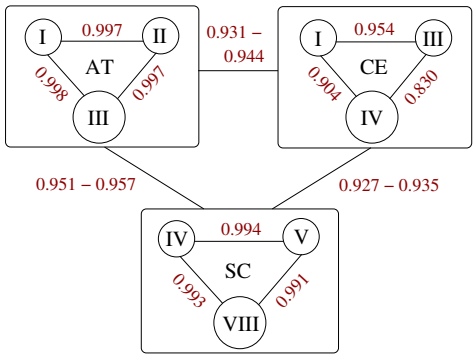


Figure 5.13: Pearson correlations between θ_3^{vdo} signatures of AT, CE, and SC chromosomes

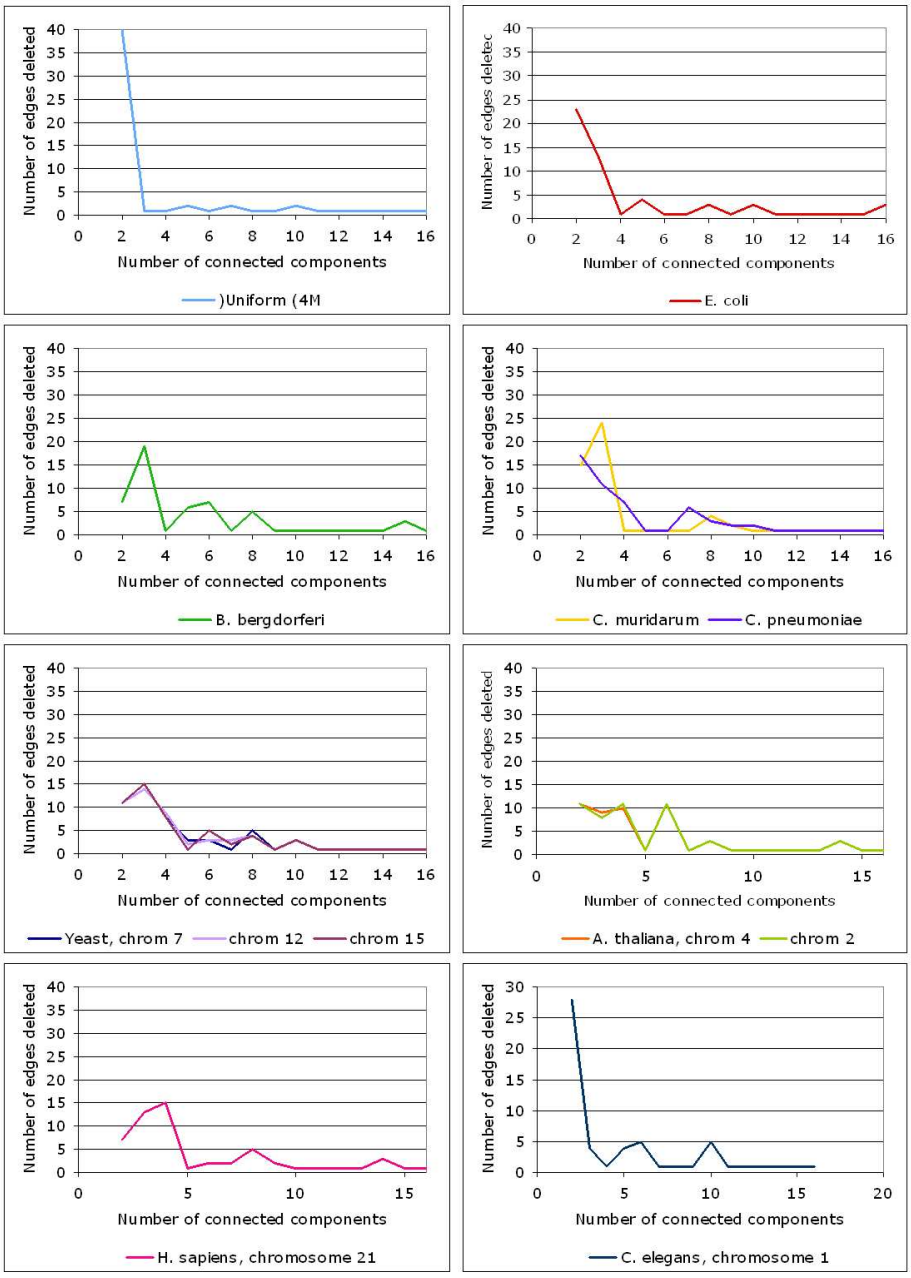


Figure 5.14: θ_2^{ced} signatures of various species. Complete chromosomal sequences have been used to generate signatures. The first figure corresponds to the θ_2^{ced} signature for a sequence generated by a DBC with uniform transition probabilities.

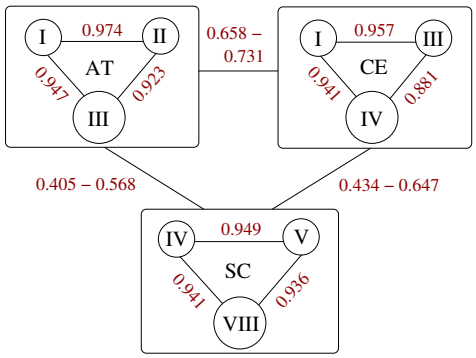


Figure 5.15: Pearson correlations between θ_3^{ced} signatures of AT, CE, and SC chromosomes.

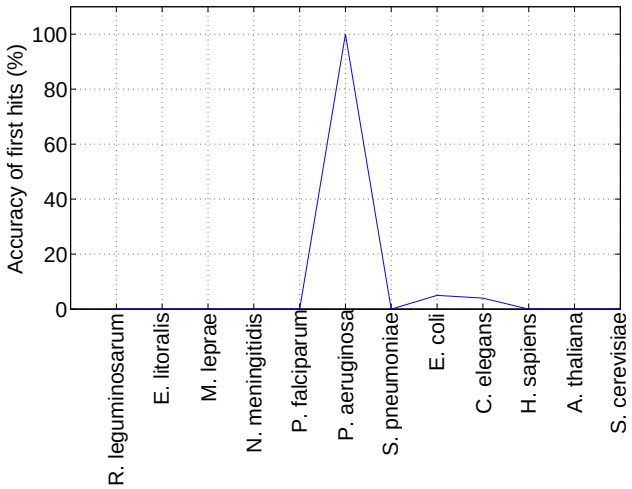
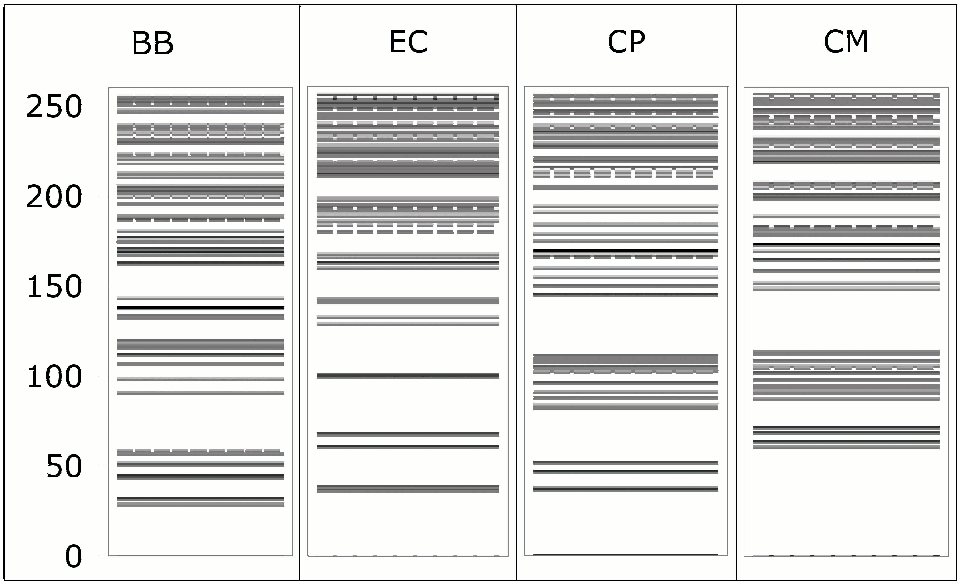


Figure 5.16: Accuracy of first hits of the θ_2^{ced} signature. Sample sequences of length 10 kb and a database of 12 diverse species have been used.

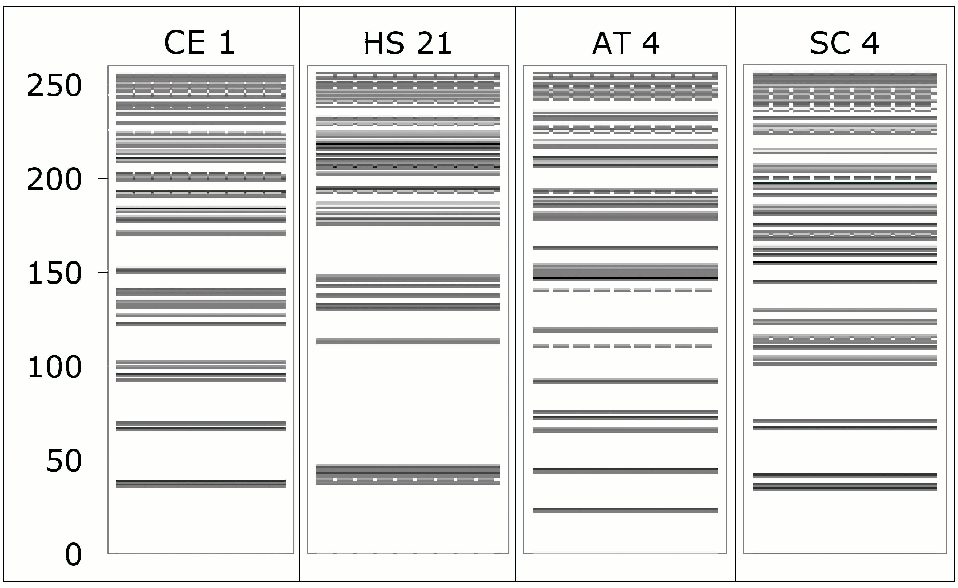
is the 4^w-vector whose i^{th} component is the total number of edge deletions required to isolate the vertex x_i, where x_i is the i^{th} element of $\mathcal{S}^w$ in lexicographic order. Figure 5.17 illustrates that the θ_3^{oed} bar code of each species is unique and sufficiently different from the θ_3^{oed} bar codes of other species.

The accuracies of origin prediction of the three signatures θ_2^{wcv}, θ_2^{dor}, and θ_2^{oed} were compared in two scenarios. The first scenario consisted of testing the accuracy of origin prediction, while choosing from an existing database consisting of signatures of far-away species. We used 12 species for this purpose as described in Table 6.1. Sequences of length 50 kb that were randomly sampled from genomic sequences of these species were used as input data. The results are shown in Figure 5.18(a).

The second scenario consisted of testing the accuracy of origin prediction while choosing from an existing database consisting of signatures of closely-related species. We used 20 species for sampling purposes as described in Table 6.2, while the database consisted of signatures of 52 α-proteobacteria as listed in Table 5.2. In this work, we have excluded plasmids from all experiments involving bacteria. Sequences of length 50 kb that were randomly sampled from genomic sequences of these species were used as input data. The results are shown in Figure 5.18(b).

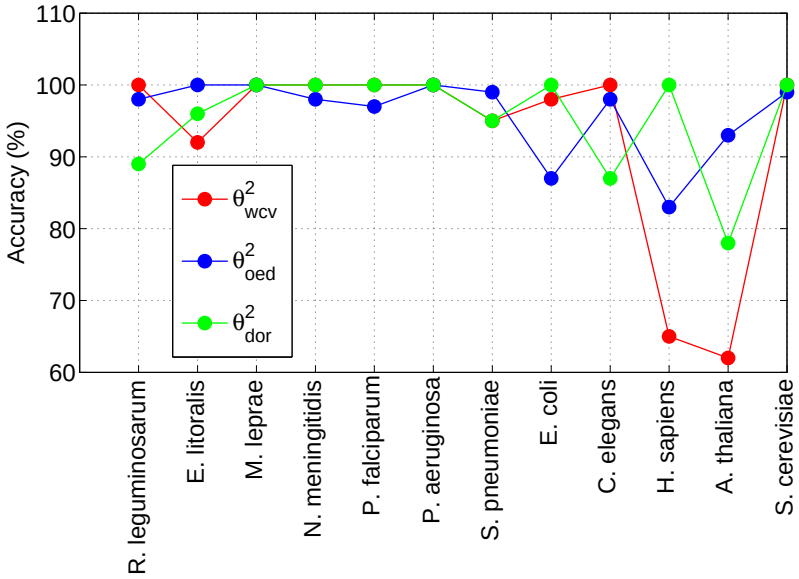


(a)

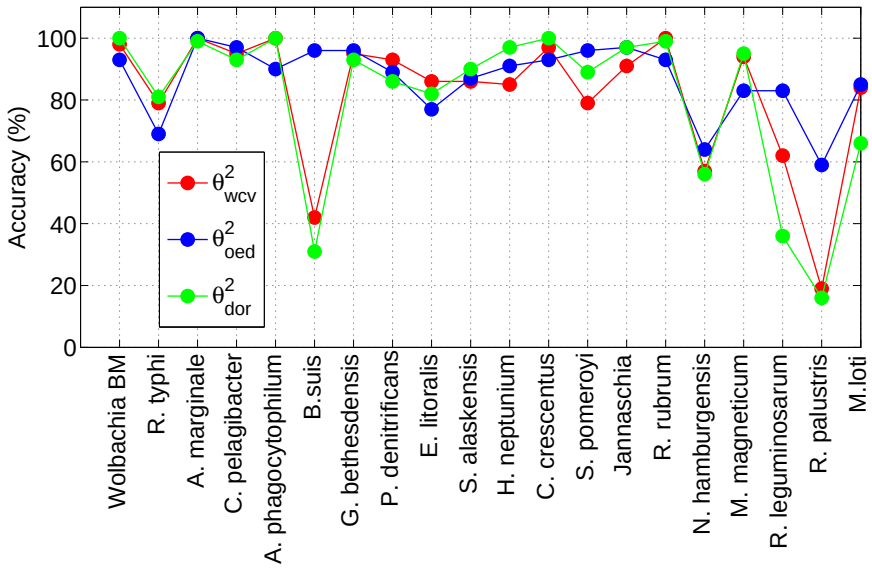


(b)

Figure 5.17: θ_3^{ed} signatures for (a) 4 prokaryotes and (b) 4 eukaryotes. Numbers denote chromosomes. The above is a gray scale representation. Each shaded-bar represents a specific component in the signature.



(a)



(b)

Figure 5.18: Comparison of accuracies of θ_2^{wcv}, θ_2^{oed}, and θ_2^{dor} signatures. We have examined origin prediction of random 50 kb segments taken from the species on the x-axis. 100 samples were used to determine the accuracy for each species. In (a), the database of 12 diverse species was used, while in (b), the database of 20 α-proteobacteria was used.

Table 5.2: List of 53 α-proteobacterial species. List of α-proteobacteria and their genomic sequences used to build a large database of closely-related species.

Species	**Size**	**NCBI accession numbers**
A. marginale	1214881	NC_004842
A. phagocytophilum	1492378	NC_007797
A. tumafaciens	2882199	NC_003062, NC_003063, NC_003304, NC_003305
B. abortus, chromosomes 1, 2	2154688	NC_006932, NC_006933
B. bacilliformis, KC583	1465745	NC_008783
B. henselae, Houston-1	1958717	NC_005956
B. japonicum, USDA-110	9235994	NC_004463
B. melitensis, 16M, chromosomes 1, 2	2147476	NC_003317, NC_003318
B. melitensis, biovar-Abortus-2308, chromosomes 1, 2	2151768	NC_007618, NC_007624
B. quintana	1604058	NC_005955
B. suis, chromosomes 1, 2	2137988	NC_004310, NC_004311
C. crescentus	4074408	NC_002696
C. pelagibacter, ubique-HTCC1062	1327544	NC_007205
E. canis, Jake	1333891	NC_007354
E. chaffeensis, Arkansas	1193136	NC_007799
E. litoralis, HTCC2594	3096085	NC_007722
E. ruminantium, Gardel	1521430	NC_006831
E. ruminantium, Welgevonden	1534678	NC_006832
G. bethesdensis, CGDNIH1	2747130	NC_008343
G. oxydans, 621H	2740851	NC_009977
H. neptunium, ATCC-15444	3758031	NC_008358
Jannaschia sp. CCS1	4379731	NC_007802
Mesorhizobium sp. BNC1	4475553	NC_008254
M. loti	7136665	NC_002678
M. magneticum, AMB-1	5038190	NC_007626
M. maris, MCS10	3416978	NC_008347
N. aromaticivorans, DSM-12444	3612554	NC_007794
N. hamburgensis, X14	4470001	NC_007964
N. winogradskyi	3450775	NC_007406

P. denitrificans, PD1222, chromosomes 1, 2	2893125	NC_008686, NC_008687
R. bellii, RML369-C	1543895	NC_007940
R. conorii, Malish7	1286962	NC_003103
R. denitrificans, OCh-114	4192225	NC_008209
R. etli	4444225	NC_007761
R. felis, URRWXCal2	1506440	NC_007109
R. leguminosarum	5129417	NC_008380
R. palustris, BisA53	5584227	NC_008435
R. palustris, BisB18	5592696	NC_007925
R. palustris, BisB5	4962694	NC_007958
R. palustris, CGA009	5537284	NC_005296
R. palustris, HaA2	5407903	NC_007778
R. prowazekii, Madrid-E	1127486	NC_000963
R. rubrum, ATCC-11170	4415090	NC_007643
R. sphaeroides, 2. 4. 1	3234254	NC_007493
R. typhi, Wilmington	1127456	NC_006142
S. alaskensis, RB2256	3393039	NC_008048
Silicibacter sp. TM1040	3246738	NC_008044
S. meliloti	1373665	NC_003037, NC_003047, NC_003078
S. pomeroyi, DSS-3	4168226	NC_003911
Wolbachia, *Brugia malayi* endosymbiont	1095613	NC_006833
Wolbachia, *Drosophila melanogaster* endosymbiont	1285992	NC_002978
Z. mobilis	2085879	NC_006526

5.8 Discussion and Conclusions

To compare the accuracies of the three signatures we used sequence samples of fairly large length, i.e., 50 kb. From Figure 5.18(a), we observe that the θ_2^{oed} signature outperforms the θ_2^{wfv} signature 4/12 times while the latter outperforms the former 4/12 times. Similarly, the θ_2^{oed} signature outperforms the θ_2^{dor}

signature 5/12 times while the θ_2^{dor} signature outperforms the θ_2^{oed} signature 5/12 times. From the plot in both Figure 5.18(b), we observe that the θ_2^{oed} signature outperforms the θ_2^{wfv} signature 11/20 times while the latter outperforms the former 8/20 times. Similarly, the θ_2^{oed} signature outperforms the θ_2^{dor} signature 10/20 times while the θ_2^{dor} signature outperforms the θ_2^{oed} signature 9/20 times. Observe also, that, while distinguishing a DNA segment sampled from *R. leguminosarum* is more accurate using a database of diverse species, the identification becomes less accurate when a database of closely-related species. While it appears that the graph-based θ_2^{oed} signature is evenly matched with both the dinucleotide frequency based signatures, it is notable that their performances are complimentary to each other. For instance, observe that, in the case of *B. suis*, the graph-based θ_2^{oed} signature performs much better than the θ_2^{dor} and θ_2^{wcv} signatures. suggesting that the sequence features captured by the θ_2^{oed} signature are more well-conserved in *B. suis* than dinucleotide counts or odds ratios. The complementarity of accuracies could be attributed to the fact that the two categories of signatures pick out entirely different features of sequences. This led to the possibility that a signature that picked out both graph structure-based features and oligonucleotide frequency based features would have substantially higher accuracy than either of these signatures alone. In further chapters, we further improve the θ^{oed} signature to obtain the more accurate θ^{ovif} signature. We also derive a stronger and more accurate word frequency based signature, the π signature. In Chapter 6, we combine these two signatures to obtain the θ^{dbc} signature that has a higher accuracy of origin prediction than any of the signatures discussed so far.

Chapter 6

The de Bruijn chain signature

In this chapter, we characterize the de Bruijn chain signature θ^{dbc}. We build a theoretical framework within which the properties of the θ_w^{dbc} signature can be explored. Within this framework, we characterize the θ_2^{dbc} signature in particular, although the methods and bounds presented for the θ_2^{dbc} signature are applicable to higher-order θ_w^{dbc} signatures. We derive and evaluate probabilistic upper bounds on the separation between the θ_2^{dbc} signatures of sequences generated by the same de Bruijn chain and probabilistic lower bounds on the separation between the θ_2^{dbc} signatures of sequences generated by different de Bruijn chains. We present results that show the accuracy of target identification by the θ_2^{dbc} signature when the database consists of far-away species as well as closely-related species. In the end, we combine the powers of the θ_2^{dbc} and θ_2^{dor} signatures for even more accurate origin predictions.

6.1 Theory and Methods

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. 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 π_{x_i}. The stationary probability for x_i, when estimated from a given sequence H, is denoted by $\pi_{x_i}(H)$.

As observed in Chapter 2, the θ_2^{dbc} signature is a concatenation of the π_2 signature and the $\theta_2^{ovif}/4$ signature.

First we develop a framework for characterizing π_2.

6.1.1 Separation between π_2 signs 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 $\pi_w(H)$. Let h be a much shorter subsequence of H with transition matrix P' and stationary distribution $\pi_w(h)$. Assuming that P' is irreducible, P' is a perturbed form of P. When P and P' are close, the distance between $\pi_w(H)$ and $\pi_w(h)$ is very small and can be bounded.

Recall that S is a genomic sequence over the alphabet Σ_{DNA}.

Solan and Vieille [116] have defined a measure of closeness of P' to P. They define ζ as

$$\zeta_P = \min_{C \subset \mathcal{S}} \sum_{s \in C} \pi_w(H) \Pr\left[s \to \bar{C}\right].$$

They state that P' is (ϵ, b)-close to P if for all pairs of states $s, t \in \mathcal{S}^w$,

$$\left|1 - \frac{P'(s \to t)}{P(s \to t)}\right| \leq b$$

whenever (a) $\pi_s^w(H) P(s \to t) \geq \epsilon\zeta_P$ or (b) $\pi_s^w(H) P'(s \to t) \geq \epsilon\zeta_P$. 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

- P' is irreducible and
- Its stationary distribution $\pi_w(h)$ satisfies

$$\left|1 - \frac{\pi_s^w(h)}{\pi_s^w(H)}\right| \leq 18bL.$$

A more detailed account of the above intuition can be found in Solan and Vieille [116].

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 $\pi_w(H)$ and closely approximated by $\pi_w(h)$. Therefore, the estimated stationary distribution of the DBC that generates a genomic sequence, can serve as a genomic signature.

Our results [53] (not shown here) suggest that $\theta_w^{wfv}(H) \approx \pi_w(H)$, and $\pi_w(h) \approx \pi_w(H)$, while $\theta_w^{wfv}(h)$ might not display such similarity to either $\theta_w^{wfv}(H)$ or $\pi_w(H)$. This property is conserved for a wide range of lengths of h (tested for ≥ 5 kb). In Theorem 6.2, 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 6.1. *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 $\pi_x(H)$ of x estimated from H. Then for $\tau > 0$ and $T = n\tau$,*

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

where

$$\mathcal{L}^{\pi}(x) \quad = \quad \exp\left(\frac{T^2}{2n\pi_x}\right)$$

and

$$\mathcal{U}^{\pi}(x) \quad = \quad \left(\frac{\mathrm{e}^{\frac{T}{n\pi_x}}}{\left(1+\frac{T}{n\pi_x}\right)^{1+\frac{T}{n\pi_x}}}\cdot\right)^{n\pi_x}.$$

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 $\mathbf{E}\left[X_x\right] \approx n\pi_x$. Now,

$$\Pr\left[|\pi_x - \mathbf{E}\left[\pi_x\right]| > \tau\right] \quad = \quad \Pr\left[|X_x - \mathbf{E}\left[X_x\right]| > n\tau\right].$$

Let $T = n\tau$. Since X_x can be expressed as a sum of independent indicator random variables, Chernoff's bounds [87] are applicable. For the lower tail of the above probability, the following Chernoff bound [87] is applicable:

$$\Pr\left[X_x < (1-\delta)\mu\right] < \mathrm{e}^{\frac{-\mu\delta^2}{2}},$$

where $\mu = \mathbf{E}\left[X_x\right]$. Using

$$\begin{aligned}\Pr\left[X_x - \mathbf{E}\left[X_x\right] < -T\right] &= \Pr\left[X_x < n\pi_x - T\right] \\ &= \Pr\left[X_x < (1-\delta)n\pi_x\right]\end{aligned}$$

we get

$$\begin{aligned}n\pi_x - T &= (1-\delta)n\pi_x \\ \text{or } \delta &= \frac{T}{n\pi_x}.\end{aligned}$$

Therefore, the lower tail probability is bound as follows:

$$\begin{aligned}\Pr\left[X_x - \mathbf{E}\left[X_x\right] < -T\right] &< \exp\left(\frac{-n\pi_x}{2} \cdot \left(\frac{T}{n\pi_x}\right)^2\right) \\ &= \exp\left(\frac{-T^2}{2n\pi_x}\right).\end{aligned}$$

For the corresponding upper tail of the probability, the following Chernoff's bound [87] is applicable:

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

Using

$$\begin{aligned}\Pr\left[X_x - \mathbf{E}\left[X_x\right] > T\right] &= \Pr\left[X_x > n\pi_x + T\right] \\ &= \Pr\left[X_x > (1+\delta)n\pi_x\right]\end{aligned}$$

we get

$$\begin{aligned}n\pi_x + T &= (1+\delta)n\pi_x \\ \text{or } \delta &= \frac{T}{n\pi_x}.\end{aligned}$$

Therefore, the upper tail probability is bound as follows:

$$\Pr\left[X_x - \mathbf{E}\left[X_x\right] > T\right] < \left(\frac{\mathrm{e}^{\frac{T}{n\pi_x}}}{\left(1+\frac{T}{n\pi_x}\right)^{1+\frac{T}{n\pi_x}}}.\right)^{n\pi_x}.$$

Combining the two tail probabilities we have

$$\begin{aligned}\Pr\left[|\hat{\pi}_x - \mathbf{E}\left[\hat{\pi}_x\right]| > \tau\right] &= \Pr\left[|X_x - \mathbf{E}\left[X_x\right]| > T\right] \\ &\leq \mathcal{L}^{\pi}(x) + \mathcal{U}^{\pi}(x).\end{aligned}$$

The lemma follows. □

Theorem 6.2. *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 random variables representing 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 $n\pi_x$. Then for $\tau > 0$ and $T = n\tau$,*

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

Proof. Using the bound for the stationary distribution of each dinucleotide as derived in Lemma 6.1 and applying the union bound we have

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

and

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

The expected value of π_x for any x is the same in both sequences H_1 and H_2. Therefore,

$$d\left((\hat{\pi}^1 - \mathbf{E}\left[\hat{\pi}^1\right]), (\hat{\pi}^2 - \mathbf{E}\left[\hat{\pi}^2\right])\right) \quad = \quad d\left(\hat{\pi}^1, \hat{\pi}^2\right).$$

Using the union bound we get,

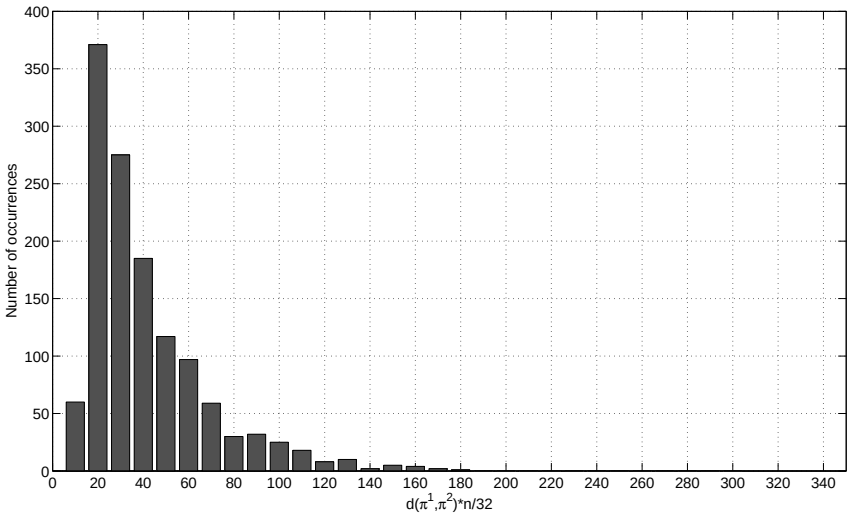
$$\begin{aligned}
\Pr\left[d\left(\hat{\pi}^1, \hat{\pi}^2\right) > 32\tau\right] &= \Pr\left[d\left(\hat{\pi}^1, \hat{\pi}^2\right) > 32T/n\right] \\
&= \Pr\left[d\left(\hat{\pi}^1 - \mathbf{E}\left[\hat{\pi}^1\right], \hat{\pi}^2 - \mathbf{E}\left[\hat{\pi}^2\right]\right) > 32T/n\right] \\
&\leq \Pr\left[|\hat{\pi}^1 - \mathbf{E}\left[\hat{\pi}^1\right]| > 16T/n\right] + \Pr\left[|\hat{\pi}^2 - \mathbf{E}\left[\hat{\pi}^2\right]| > 16T/n\right].
\end{aligned}$$

The theorem follows. □

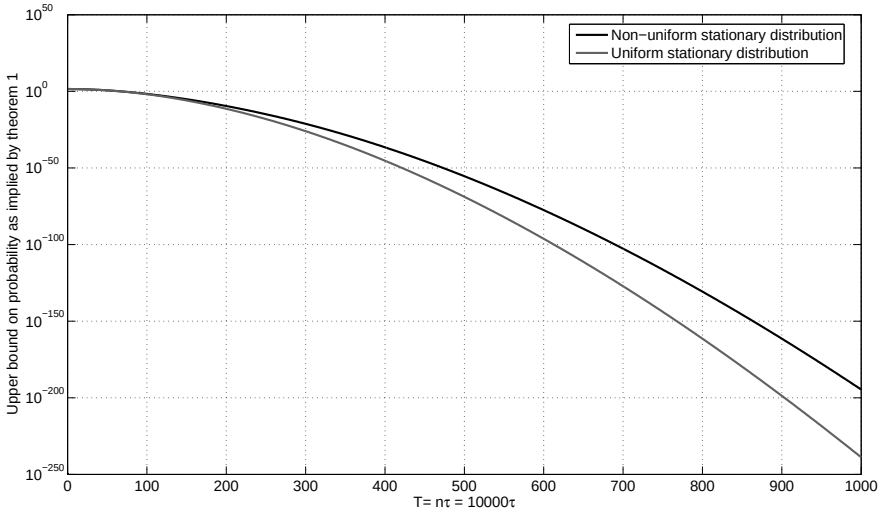
We study the nature of the above bound in Theorem 6.2 as follows. In Figure 6.1(a), we have plotted the distribution of $T = n\tau$ values using τ values computed from L_1 distances between sequences sampled from the same organism. Sequences of size 10 kilobases were used. A set of genomic sequences were randomly selected. From each genomic sequence 100 pairs of sub-sequences were independently sampled at random, their stationary distributions were estimated, and the L_1 distance between each pair of stationary distributions was recorded. Both τ and T values were computed from this distance and their distribution plotted as in Figure 6.1(a). In Figure 6.1(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. Note that approximately for $T > 160$, the corresponding probability of separation is very low. This illustrates that the bound displays a strong synergy to real data from genomes.

6.1.2 Separation between θ_2^{ovif} signs derived from sequences generated by the same DBC

Next, we present bounds for separation between the θ_2^{ovif} signatures derived from a pair of genomic sequences generated by the same DBC. We begin by characterizing the distribution of the transition probability between two states. Let the transition $t : \sigma_1 \cdots \sigma_w \rightarrow \sigma_2 \cdots \sigma_{w+1}$ be defined. Let X and Y be random variables



(a)



(b)

Figure 6.1: Plot of upper bounds derived in Theorem 6.2. (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 6.2.

denoting the number of occurrences of $\alpha = \sigma_1\sigma_2\cdots\sigma_{w+1}$ and $\beta = \sigma_1\cdots\sigma_w$ respectively 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 6.3 presents an upper bound on the probability of a specified separation between the frequency of a given transition t and its expected value.

Lemma 6.3. *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 $\tau > 0$,*

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

where,

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

and

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

Proof. Recall that the θ_2^{ovif} signature is scaled by 4 to maintain similar orders of magnitude as the π_w signature (Section 6.1). Let X_α and X_β be random variables representing the number of occurrences of strings α and β respectively in H.

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

Since X_α can be represented as a sum of independent indicator random variables with $\mathbf{E}\left[X_\alpha\right] = c\pi_\alpha/\pi_\beta$, Chernoff's bounds [87] are applicable to the probability

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

Consider the lower tail probability

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

Chernoff's lower tail bounds [87] are of the form

$$\Pr\left[X < (1-\delta)\mu\right] < \mathrm{e}^{\frac{-\mu\delta^2}{2}},$$

where $\mu = \mathbf{E}\left[X_\alpha\right]$. Using

$$\frac{c\pi_\alpha}{\pi_\beta} - 4\tau c \quad = \quad (1-\delta)\frac{c\pi_\alpha}{\pi_\beta}$$

we get

$$\delta \quad = \quad \frac{4\tau\pi_\beta}{\pi_\alpha}.$$

Therefore, the lower tail probability is bounded as follows:

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

Now consider the upper tail probability

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

The following Chernoff's bound [87] is applicable.

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

Using

$$\frac{c\pi_\alpha}{\pi_\beta} + 4\tau c \quad = \quad (1+\delta)\frac{c\pi_\alpha}{\pi_\beta}$$

we get

$$\delta \quad = \quad \frac{4\tau\pi_\beta}{\pi_\alpha}.$$

Therefore, the upper tail probability is bounded as follows:

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

Combining the two tail probabilities we get

$$\begin{aligned}
&\Pr\left[|Z/4 - \mathbf{E}\left[Z/4\right]| \geq \tau\right] \\
&< \sum_{c=1}^{\infty} \left(\exp\left(-8c\tau^2 \frac{\pi_\beta}{\pi_\alpha}\right) + \left(\frac{\mathrm{e}^{\frac{4\tau\pi_\beta}{\pi_\alpha}}}{\left(1+\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^{1+\frac{4\tau\pi_\beta}{\pi_\alpha}}} \right)^{\frac{c\pi_\alpha}{\pi_\beta}} \right) \frac{\mathrm{e}^{-n\pi_\beta}(n\pi_\beta)^c}{c!} \\
= \quad & \mathrm{e}^{-n\pi_\beta} \left(\sum_{c=1}^{\infty} \frac{\left(\exp\left(-8\tau^2\frac{\pi_\beta}{\pi_\alpha}\right)(n\pi_\beta)\right)^c}{c!} + \sum_{c=1}^{\infty} \frac{\left(\left(\frac{\mathrm{e}^{\frac{4\tau\pi_\beta}{\pi_\alpha}}}{\left(1+\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^{1+\frac{4\tau\pi_\beta}{\pi_\alpha}}} \right)^{\frac{\pi_\alpha}{\pi_\beta}} (n\pi_\beta) \right)^c}{c!} \right) \\
= \quad & \mathrm{e}^{-n\pi_\beta} \left(\exp\left(\exp\left(-8\tau^2\frac{\pi_\beta}{\pi_\alpha}\right)(n\pi_\beta)\right) + \exp\left(\left(\frac{\mathrm{e}^{\frac{4\tau\pi_\beta}{\pi_\alpha}}}{\left(1+\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^{1+\frac{4\tau\pi_\beta}{\pi_\alpha}}} \right)^{\frac{\pi_\alpha}{\pi_\beta}} (n\pi_\beta) \right) - 2 \right).
\end{aligned}$$

The lemma follows. □

We assume the existence of a maximum transition probability among all probabilities associated with transitions to or from any given state in Assumption 6.1.

Assumption 6.1. *Consider an order-2 DBC $\mathcal{DC}$ that generates sequence H. Let $\hat{\mathcal{DC}}$ be the DBC reconstructed from H. Given a state $\beta \in \Sigma_{\mathrm{DNA}}^w$ 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 $\sigma \in \Sigma_{\mathrm{DNA}}$. There exists a positive constant s, and a maximum transition $t^* \in$ trans(β) in $\mathcal{DC}$ such that, for all $t \in$ trans(β) $\setminus \{t^*\}$,*

$$p(t^*) - p(t) > s,$$

where $p(t)$ denotes the probability associated with the transition t. For some ς, where $0 < \varsigma \leq 1$, and transitions $t \in$ trans(β), the probability that the same transition $t^* \in$ trans(β) 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 $\beta \in \Sigma_{\mathrm{DNA}}^w$, 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 $\sigma \in \Sigma_{\mathrm{DNA}}$. 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 $\sigma \in \Sigma_{\mathrm{DNA}}$.

The L_1 distance between the θ_2^{ovif} signatures of sequences generated by the same DBC is bounded in Theorem 6.4.

Theorem 6.4. *Assume Assumption 6.1. 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 6.3, Assumption 6.1, and the union bound we get

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

and

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

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

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

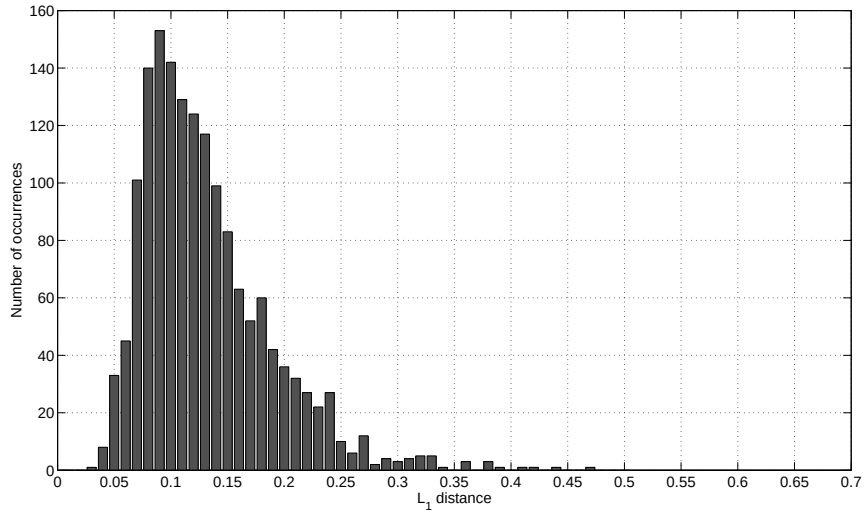
We get

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

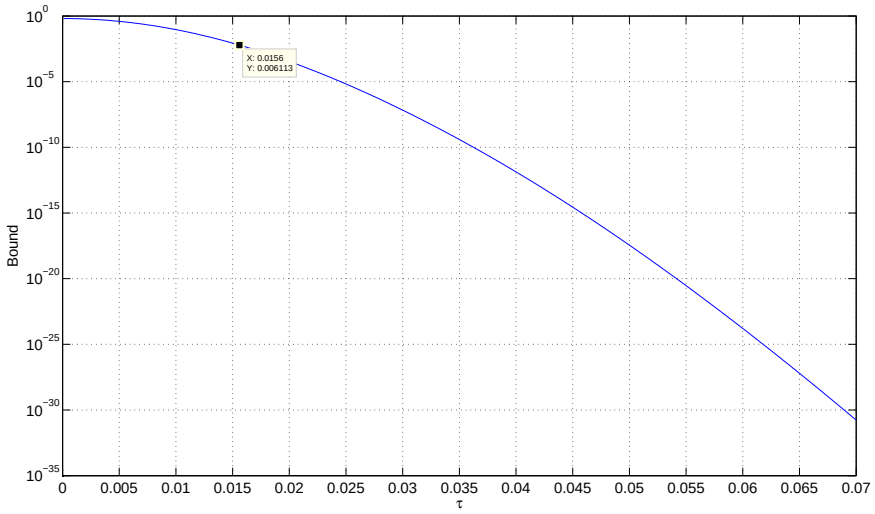
The theorem follows. □

We now analyze the behavior of the upper bound in Theorem 6.4 when applied to real data. For a randomly selected set of genomic sequences, 100 pairs of sequences of length 100 kilobases 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 6.2(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 6.2(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.



(a)



(b)

Figure 6.2: Plot of upper bounds derived in Theorem 6.4. (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 6.4. 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).

6.1.3 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 6.5 proves that the separation between their θ_w^{dbc} signatures is less than a specified threshold with high probability.

Theorem 6.5. *Let $\mathcal{DC}$ be an order 2 DBC with underlying stationary distribution π. 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 DBC signatures. Similarly, let $\hat{\pi}^1$ and $\hat{\pi}^2$ be their estimated order-2 stationary distributions and θ_1^{ovif} and θ_2^{ovif} be their order-2 OVIF signatures, respectively. Then,*

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

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

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

The theorem follows using the results from Theorems 6.2 and 6.4. □

6.1.4 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 the stationary distributions estimated from the respective two sequences and θ_1^{ovif} and θ_2^{ovif} be their respective OVIF signatures.

Then, the distance $d\left(\theta_1^{dbc},\theta_2^{dbc}\right)$ can distinguish $\mathcal{DC}_1$ and $\mathcal{DC}_2$. Assumptions 6.2 formalizes the separation of estimated stationary distributions of genomic sequences obtained from different organisms, while Assumption 6.3 formalizes the probability of the maximum transition being different for a given state using genomic sequences obtained from different organisms.

Assumption 6.2.

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

Assumption 6.3.

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

Theorem 6.6. *If there exist constants γ and ν as in Assumptions 6.2 and 6.3 then,*

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

Proof. Treating $d\left(\theta_1^{dbc},\theta_2^{dbc}\right)$, $d\left(\theta_1^{dbc},\mathbf{E}\left[\theta_1^{dbc}\right]\right)$, $d\left(\theta_2^{dbc},\mathbf{E}\left[\theta_2^{dbc}\right]\right)$, and $d\left(\mathbf{E}\left[\theta_1^{dbc}\right],\mathbf{E}\left[\theta_2^{dbc}\right]\right)$ as distances d_1, d_2, d_3, and d_4 respectively, in 1-dimensional space we obtain,

$$\begin{aligned} d_3 &\leq d_1+d_2+d_3 \\ \Pr\left[d_4\geq 6\cdot 16\tau\right] &\leq \Pr\left[d\geq 2\cdot 16\tau\right]+\Pr\left[d_1\geq 2\cdot 16\tau\right]+\Pr\left[d_2\geq 2\cdot 16\tau\right]. \end{aligned}$$

From Assumptions 6.2 and 6.3 we obtain,

$$d\left(\mathbf{E}\left[\theta_1^{dbc}\right],\mathbf{E}\left[\theta_2^{dbc}\right]\right)\geq 6\cdot 16\tau.$$

We have,

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

The theorem follows. □

We demonstrate Assumptions 6.2 and 6.3 using sequences from the species *C. elegans* and *P. falciparum*. Figure 6.3 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 π^1 and π^2 is 0.4735. From Assumption 6.2 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 6.3. A similar scenario is observed for Assumption 6.3. 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.

In Theorem 6.6, each negative term in the R.H.S. is very small, making the total probability on the R.H.S a very large value. Theorem 6.6 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.

6.1.5 Algorithm

Let H be a genomic sequence whose origin is unknown. Algorithm 1 is used to approximate the origin of an unknown sequence using order-2 θ^{dbc} signatures. For an available genomic sequence H, the corresponding DBC signature $\theta_2^{dbc}(H)$ is first computed. This signature is then compared with all DBC signatures of order

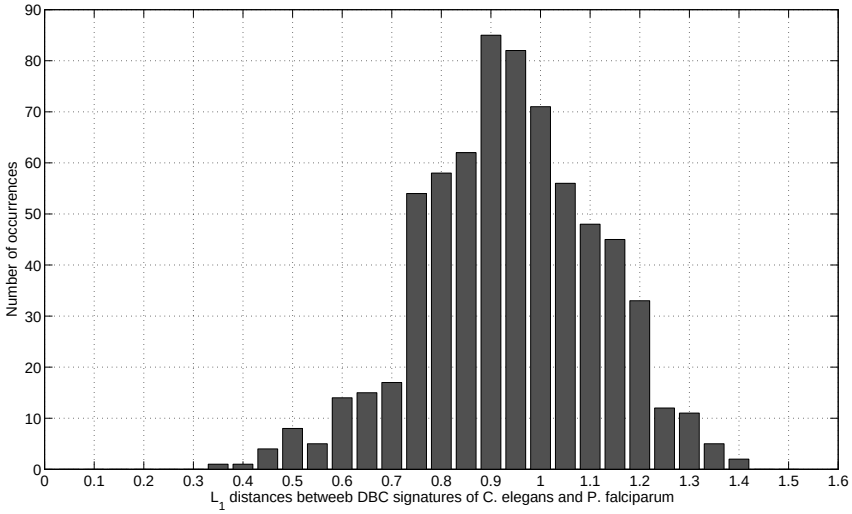


Figure 6.3: Distribution of L_1 distances between θ_2^{dbc} signatures of CE and PF. Distribution of L_1 distances between θ_2^{dbc} signatures of pairs of 10 kilobase long sequences randomly sampled from the two species *C. elegans* and *P. falciparum*.

2 stored in the database $\mathcal{D}$ using the Pearson correlation coefficient. The genome whose signature displays maximum correlation with $\theta_2^{dbc}(H)$ is predicted as the target for H.

The θ_2^{dbc} for a sequence of length n can be computed in $O(n + 16 \log n + 4096) = O(n)$ time and space. In general, the complexity of the order-w θ_w^{dbc} signature for a sequence of length n is $O(n + 4^w \log n + (4^{3w}) = O(n+64^w)$. The 4^{3w} 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.

6.2 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 of diverse genomic sequences of various lengths including α-proteobacteria, infectious bacteria, and eukaryotes. Table 6.1 displays these genomic sequences and the acronyms used for them in this chapter.

Second, we collected a set of 52 α-proteobacterial genomes including multiple strains of several species to

Algorithm 1: MATCH

INPUT: Set $\mathcal{S}$ of genomic sequences, Database $\mathcal{D}$ of existing θ_2^{dbc}s for sequences in $\mathcal{S}$, Sequence H of unknown origin.

Compute $\theta_2^{ovif}(H)$
Compute $\pi_2(H)$
$\theta_2^{dbc}(H) \leftarrow \pi_2(H) \cdot \theta_2^{ovif}(H)/4$
maxcorr $= 0$
origin$(H) = \lambda$
for each sequence $X \in \mathcal{S}$ **do**
    $\theta_2^{dbc}(X) \leftarrow \mathcal{D}(X)$
    $\rho \leftarrow \mathrm{R}(\theta_2^{dbc}(H), \theta_2^{dbc}(X))$
    if $\rho >$ maxcorr **then**
        maxcorr $\leftarrow \rho$
        origin$(H) \leftarrow$ origin(X)
    end if
end for
return origin(H)

Table 6.1: 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

Table 6.2: List of genomic sequences in the set of closely-related α-proteobacterial species

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

build a collection of genomic sequences derived from closely related species. As before, plasmid sequences were excluded. Of these 52 sequences, the 20 that were used to randomly sample shorter sequences for origin prediction are listed in Table 6.2.

Two databases of θ^{dbc} signatures were constructed; the first database $\mathcal{D}_1^{dbc}$ consisted of the signatures corresponding to the complete sequences in list L_1, while the second database $\mathcal{D}_2^{dbc}$ consisted of the signatures corresponding to the complete sequences of the 52 α-proteobacteria, *E. coli*, and the 4 higher eukaryotes from list 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.

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

First, the ability of the θ^{dbc} signature to distinguish between genomic sequences taken from distant species was tested. In the associated experiment, the variables are the order w, the sample sequence length, and the database used for matching signatures. We used orders 2, 3, 4, and 5, sequence length 10 kb, 25 kb, 50 kb, and 100 kb, and database $\mathcal{D}_1^{dbc}$ for this purpose. For each $\langle \text{order}, \text{length} \rangle$ combination, 100 samples were randomly sampled from each organism in list L_1 (Table 6.1). We do not ensure that sampled regions are non-overlapping. 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. Recall that the actual species from which the sample is taken is called the origin of the sample. In a *first hit scenario*, the origin is ranked 1. In a *good hit scenario*, the origin is ranked 1, 2, or 3. Depending on the scenario under consideration, the number of first hits (or good hits) per 100 samples is the *accuracy*.

Figure 6.4 illustrates the accuracy of first-hits for each organism in list L_1 for all points in the above experiment. For fixed order, observe that the accuracy of origin prediction increases with increasing sample size, reaching 100% first hits at length 100 K for all species at order 4. This is intuitive because a larger sequence encodes more information about the underlying DBC. This in turn leads to the calculation of a θ^{dbc} signature highly representative of the origin. The figure also suggests that the θ^{dbc} signature is more highly conserved at order 4 than at other orders. This coincides with the hypothesis behind the application TETRA [123], which also attempts to discover the origin of unknown DNA sequences, but does not work well with short sequences. However, we also 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. Intuitively, a short sequence contains maximum information about the underlying DBC of order 1. Although the corresponding θ_1^{dbc} signature can be computed in less time than higher order signatures, it encodes information about mononucleotides only, which is insufficient to accurately predict the origin of an unknown sequence. Therefore, for identification of the origin of short DNA sequences, we use the more accurate and origin-representative, but expeditiously computable order-2 θ_2^{dbc} signatures.

In Figure 6.5, we explore 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

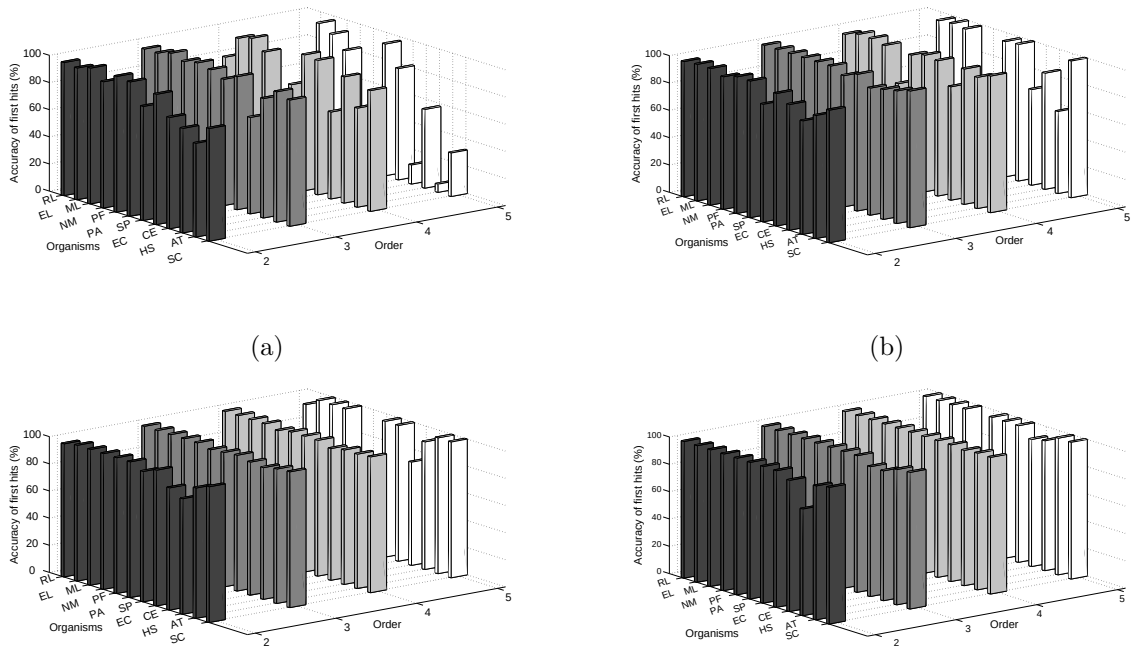


Figure 6.4: Plot of accuracies of θ^{dbc}s of orders 2 through 5. Prediction of origins from database $\mathcal{D}_1^{dbc}$ has been examined. The x-axis indicates species from list L_1. The y-axis indicates the order w. The z-axis represents the accuracy of first hits. Sample sequences of length (a) 10 kilobases, (b) 25 kilobases, (c) 50 kilobases, and (d) 100 kilobases, have been used.

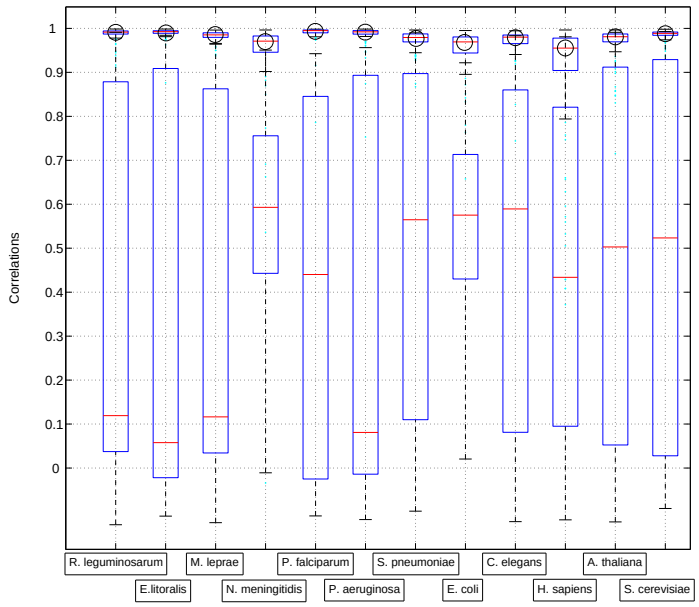
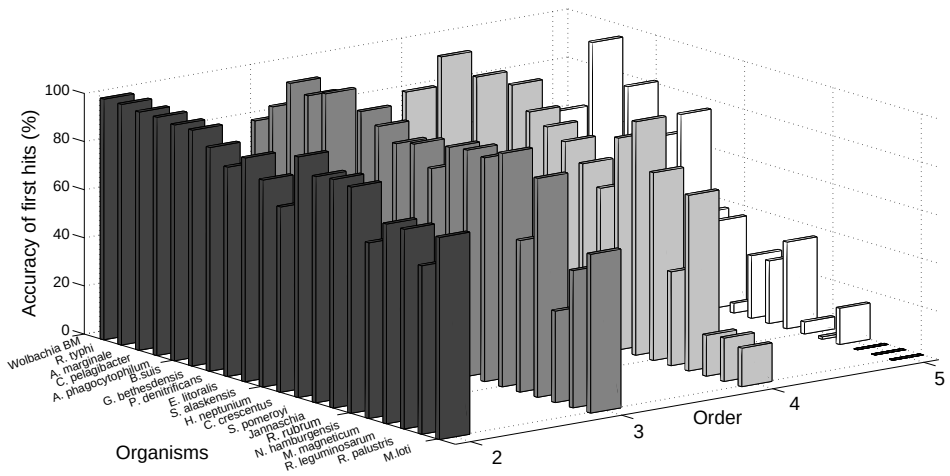


Figure 6.5: Accuracy of θ_2^{dbc}. The 12 species are on the x-axis. The small box and whisker plots near the top (with associated circles) represent the distribution of correlations of θ_2^{dbc}s of the 100 samples with the θ_2^{dbc} of their origin. The larger box and whisker plots represent the distribution of correlations with θ_2^{dbc}s of other genomes.

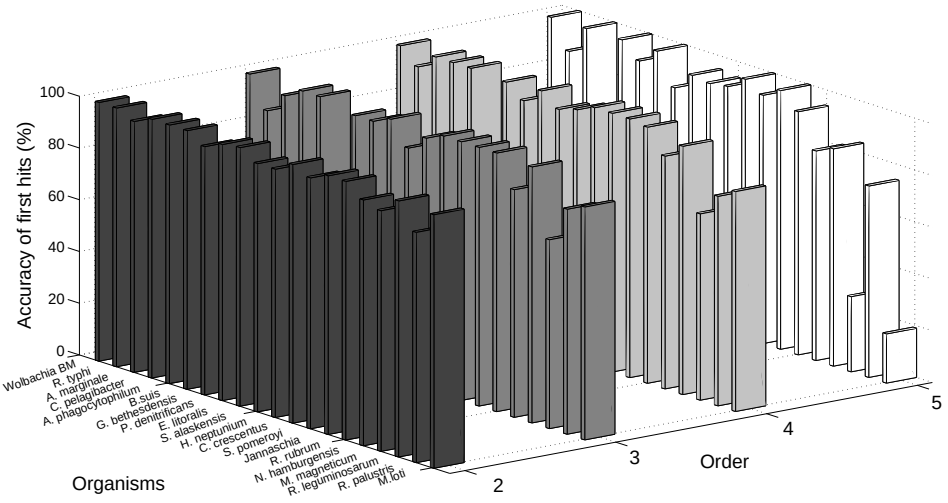
50 kb each are randomly sampled from the genome of each species. The correlation of the θ_2^{dbc} signature of each sample with the θ_2^{dbc} signature of its origin is binned separately from its correlations with the θ_2^{dbc} signatures of all other organisms. The distribution of numbers in each bin 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.

Second, the ability of the θ^{dbc} signature to distinguish between genomic sequences from closely-related species and different strains of the same species was tested. The same steps as above were followed with the following difference. Only the 20 sequences listed in list L_2 (Table 6.2) were used as sources for sampling while using $\mathcal{D}_2$ (57 signatures) for origin prediction. Similar results for the set of α-proteobacteria are presented for sample sequences of lengths 10 Kb and 50 Kb in Figure 6.6. Observe 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 the θ_2^{dbc} signature for both test cases is summarized in Figure 6.7.

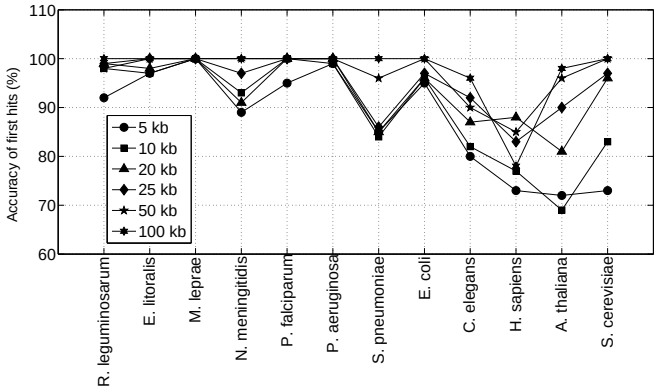


(a)

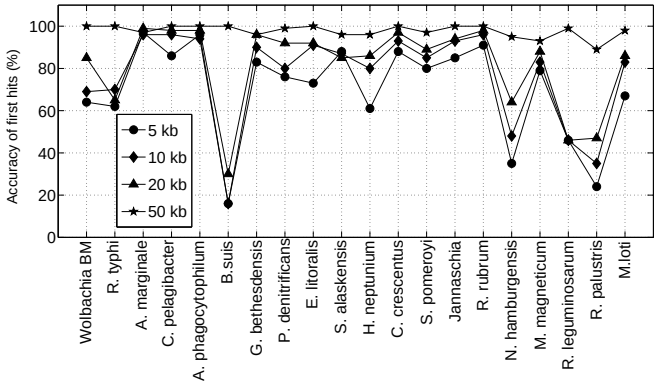


(b)

Figure 6.6: Plot of prediction accuracy vs. order for θ^{dbc} signatures. Plot of accuracies of origin prediction for orders 2, 3, 4, and 5 θ^{dbc} signatures using database $\mathcal{D}_2$ of closely related species. 100 sequences samples of lengths (a) 10 kilobases and (b) 50 kilobases were used from the α-proteobacteria in List L_2.



(a)



(b)

	Median accuracy of first hits	
Sample sequence length	**List L_1**	**List L_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	-

(c)

Figure 6.7: Summary of accuracy of first hits of θ_2^{dbc}. (a) Species in list L_1, (b) Species in list 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.

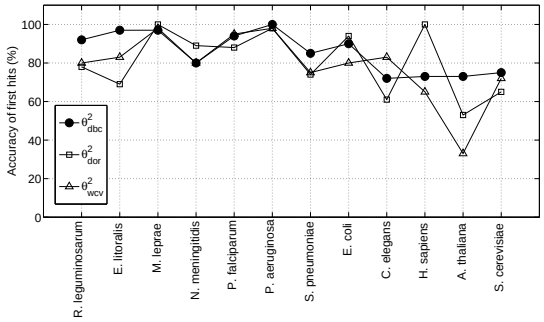
In case of list L_1, a median accuracy greater than 90% is achieved even for sequences as short as 5 kb. The median accuracy increases steadily with sample size and is 100% at a sample length of 50 kb. In the case of the human genome, the θ_2^{dbc} signature consistently does not perform well. This issue is addressed in Section 6.2.2 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 in order to be able to tell them apart. Therefore, the reduced accuracy in case of list L_2 is expected. In case of list L_2, 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.

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

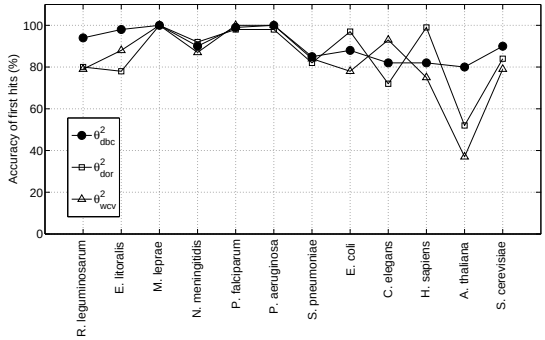
We compared the accuracy of the three signatures θ^{dbc}, θ^{dor}, and θ^{wcv} in predicting the origin of short DNA segments. The same methods and terminologies as described in Section 6.2.1 have been used. Order 2 signatures were used for several reasons. In Section 6.2.1 we found order-2 DBCs to be most representative of the origin in the case of short sequences and the corresponding θ_2^{dbc} 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 list L_1 for sampling and $\mathcal{D}_1^{dbc}$, $\mathcal{D}_1^{dor}$, and $\mathcal{D}_1^{wcv}$ databases for origin prediction. Shorter sequence samples of lengths 5 kb, 10 kb, and 20 kb were used. Figure 6.8(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.

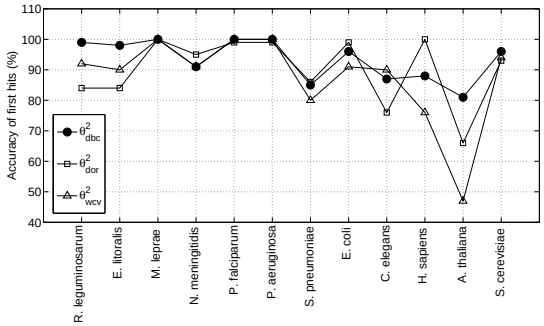
Observe that the θ_2^{dbc} signature outperforms the θ^{dor} signature for all sequence lengths by demonstrating better accuracy in the case of 8/12, 9/12, and 8/12 species for sequence lengths 5 kb, 10 kb, and 20 kb, respectively. The θ_2^{dbc} signature also outperforms the θ_2^{wcv} signature for 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 *N. meningitidis*, *E. coli*, and *H. sapiens*. Of particular interest is the human 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



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

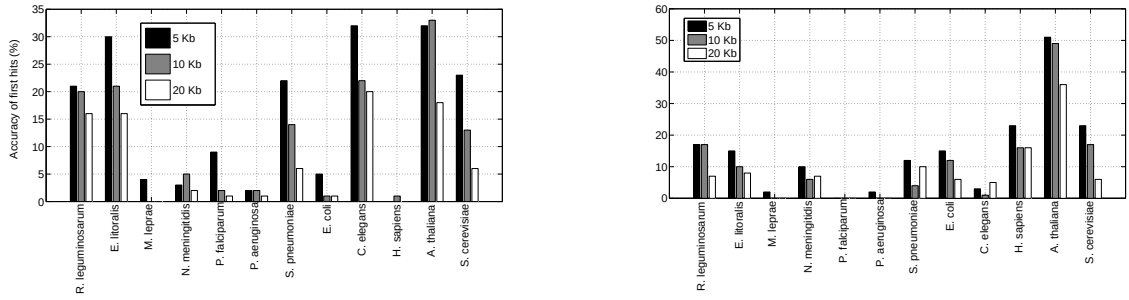


(b) 10 kb sample sequences from L_1 matched against $\mathcal{D}_1^*$



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

Figure 6.8: 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 list L_1 of diverse species 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.



(a) θ_2^{dbc} vs. θ^{dor}

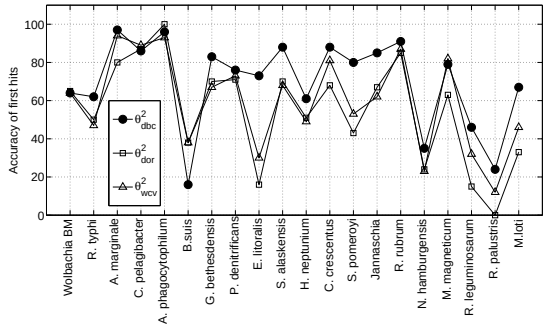
(b) θ_2^{dbc} vs. θ_2^{wcv}

Figure 6.9: Comparison of relative accuracies of θ_2^{dbc}, θ_2^{dor} and θ_2^{wcv}. 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.

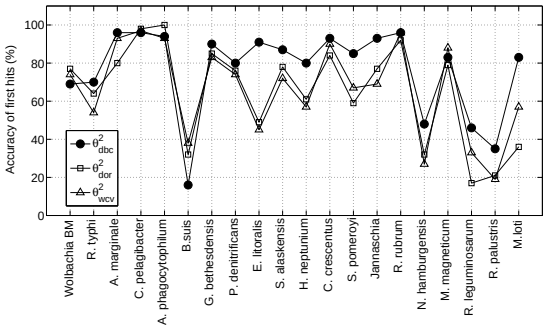
better conserved than the θ^{dor} signature. Compared with the θ_2^{wcv} signature, the θ_2^{dbc} signature consistently performs worse only in case of *C. elegans*. For all other species, the accuracy of the θ_2^{dbc} signature is better than or equal to that of the θ_2^{wcv} signature. Consider Figure 6.9. In Figure 6.9(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 6.9(b) with *C. elegans* being the only exception.

Next, we compared the abilities of the three signatures to distinguish between closely-related species while using list L_2 for sampling and the $\mathcal{D}_2^{dbc}$, $\mathcal{D}_2^{dor}$, and $\mathcal{D}_2^{wcv}$ databases for origin prediction. Short sequence samples of lengths 5 kb, 10 kb, and 20 kb were used. Figure 6.10(a), (b), and (c) illustrate the results. The same method was followed as in the previous case of diverse species. 100 subsequences were randomly sampled from each species of the 20 closely-related α-proteobacterial species on the x-axis. All three signatures were computed using each sample and correlated to their respective $\mathcal{D}_2$ databases of signatures. The accuracy of first hits are recorded on the y-axis.

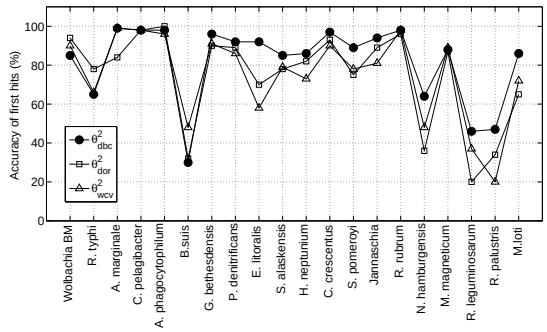
The database, in this case, contains 52 species from the same family (α-proteobacteria) and 5 other diverse species. Figure 6.10(a), (b), and (c) illustrate that the θ_2^{dbc} signature outperforms both θ^{dor} and θ_2^{wcv} signatures in the case of all sequence lengths with better predictive accuracy for 15/20 species against the



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

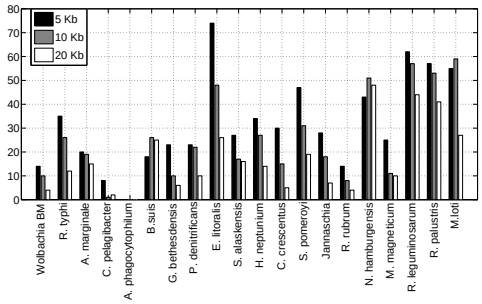


(e) 10 kb sample sequences from L_2 matched against $\mathcal{D}_2^*$

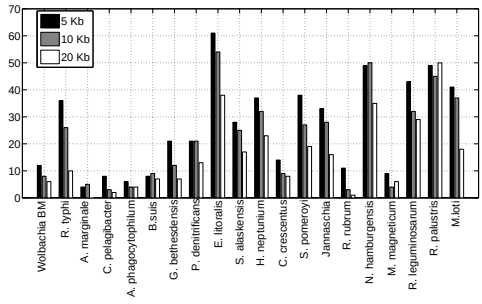


(f) 20 kb sample sequences from L_2 matched against $\mathcal{D}_2^*$

Figure 6.10: 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 list L_2 of closely related α-proteobacterial species 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.



(a) θ_2^{dbc} vs. θ^{dor}



(b) θ_2^{dbc} vs. θ_2^{wcv}

Figure 6.11: Comparison of relative accuracies of θ_2^{dbc}, θ_2^{dor} and θ_2^{wcv} for APB. 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 α-proteobacteria. 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.

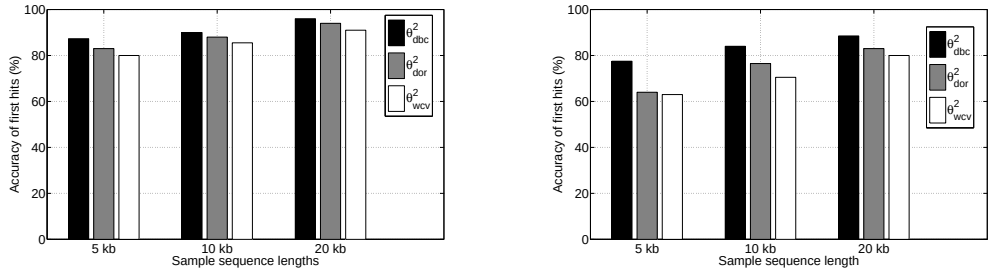
θ^{dor} signature and an average better accuracy of 16.33/20 species against the θ_2^{wcv} signature. The θ^{dor} signature appears consistently more well-conserved than the θ_2^{dbc} signature in the case of *Wolbachia*. In the comparison between the θ_2^{dbc} and θ_2^{wcv} signatures, the θ_2^{dbc} 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 sequence length as is demonstrated by the data in Figure 6.11.

For the order-2 signatures above, Figure 6.12 summarizes the median accuracy of prediction of first hits in the case of both list 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.

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

In Section 6.2.2, we demonstrated that in predicting the origin of an unknown DNA sequence the θ_2^{dbc} signature has greater accuracy than the 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 the different aspects of construction of the genomic sequence a genome that are preserved within the genome itself, while differing from those aspects in the genomic sequences of other species. So far, we have been successful in discovering some such aspects through the θ^{dbc} signature.

To test whether an even greater accuracy of origin prediction for short sequences can be achieved, we con-



(a) List L_1 (b) List L_2

Figure 6.12: Comparison of median accuracies of 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.

ducted experiments where we combined the strengths of 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, produces different results.

To reiterate our observations from Section 6.2.1, 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. Its accuracy drops only when sample sizes drop to lower lengths than 20 kb. We are interested in coupling the properties of the θ_2^{dbc} signature and the θ^{dor} signature to improve the accuracy of origin prediction in such cases. Using the product of the Pearson correlation distance and the normalized δ-distance 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 6.13. The same method as described in the previous sections was used to determine accuracy.

However, the same method does not demonstrate substantially higher accuracy than the θ_2^{dbc} signature in differentiating between closely-related species. We make this observation based on the results in in Figure 6.14. 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.

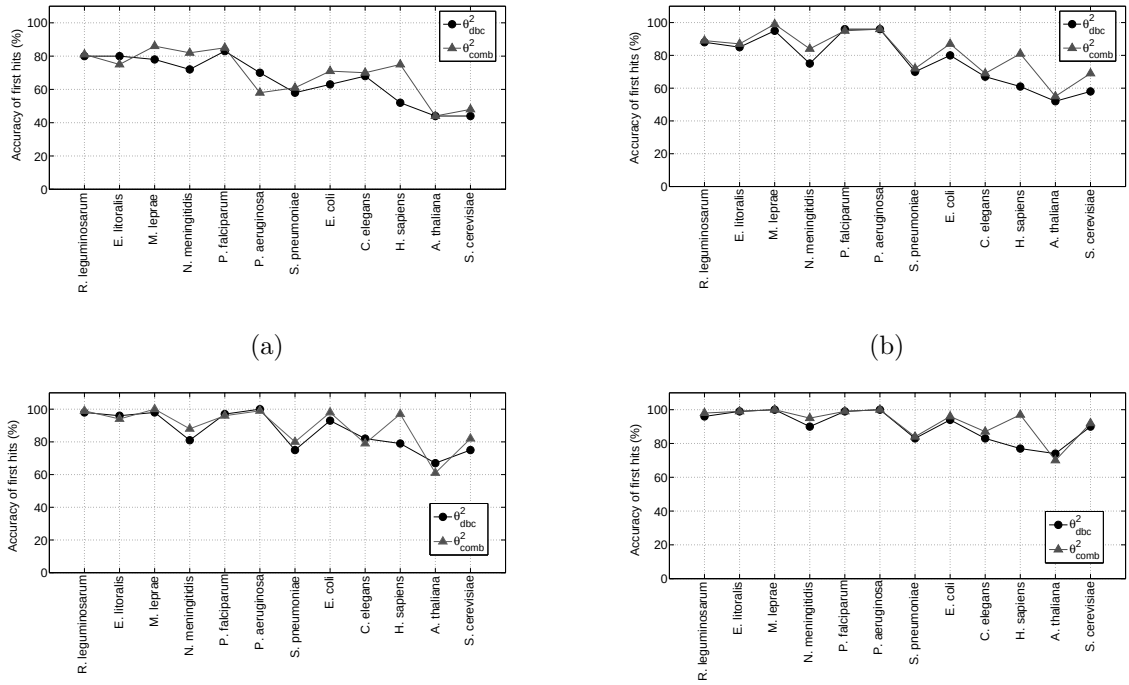


Figure 6.13: Accuracy of the combination of θ_2^{dbc} and θ^{dor} signatures. Comparison of the accuracies of the θ_2^{dbc} signature and the combination signature θ_2^{comb} of θ_2^{dbc} and θ^{dor} in predicting origins of unknown short sequences from list L_1. Sequences of lengths (a) 1 kb, (b) 2 kb, (c) 5 kb, and (d) 10 kb have been used.

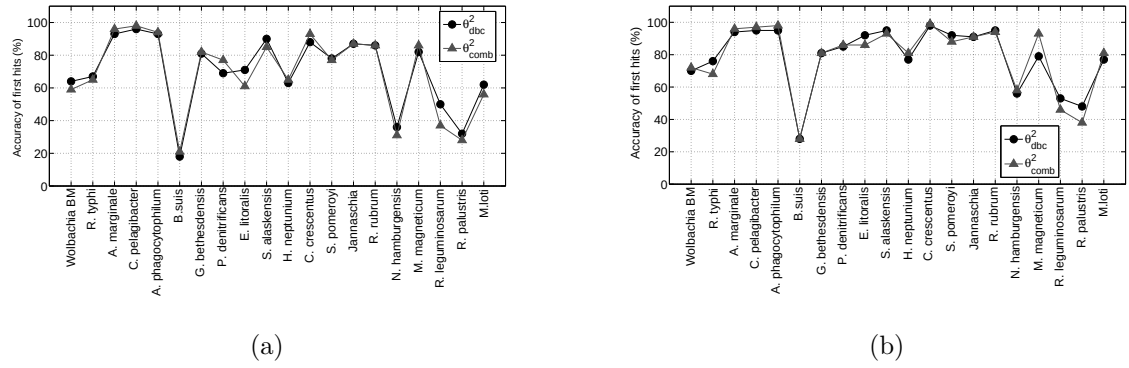


Figure 6.14: Accuracy of the combination of θ_2^{dbc} and θ^{dor} signatures for α-proteobacteria. Comparison of the accuracies of the θ_2^{dbc} signature and the combination signature θ_2^{comb} of θ_2^{dbc} and θ^{dor} in predicting origins of unknown short sequences from list L_2. Sequences of lengths (a) 5 kb and (b) 10 kb have been used.

6.2.4 Accuracies of θ_2^{dbc}, θ_2^{dor}, θ_2^{wcv}, and θ_2^{combo} for a large database of diverse species

In this section, we construct a larger diverse database of signatures for 50 diverse species while ensuring that as many regions as possible in the taxonomic tree in NCBI's database have representation in this database. In Table 6.3, we list the genomic sequences of these 50 species that we use along with their accession numbers, sizes, and positions in the taxonomic tree. We use a collection of 10 archaeal genomic sequences, 20 bacterial genomic sequences, and 20 eukaryotic genomic sequences.

Table 6.3: List of 50 diverse species taken uniformly from the taxonomic tree. List of organisms and their genomic sequences used to build a larger database of 50 diverse species while sampling species uniformly and manually from the taxonomic tree.

Aeropyrum pernix K1, complete genome	NC_000854	1693618	cellular organisms; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Aeropyrum; Aeropyrum pernix
Sulfolobus tokodaii str. 7, complete genome	NC_003106	2733328	cellular organisms; Archaea; Crenarchaeota; Thermoprotei; Sulfolobales; Sulfolobaceae; Sulfolobus; Sulfolobus tokodaii
Pyrobaculum aerophilum str. IM2, complete genome	NC_003364	2254259	cellular organisms; Archaea; Crenarchaeota; Thermoprotei; Thermoproteales; Thermoproteaceae; Pyrobaculum; Pyrobaculum aerophilum
Archaeoglobus fulgidus DSM 4304, complete genome	NC_000917	2209600	cellular organisms; Archaea; Euryarchaeota; Archaeoglobi; Archaeoglobales; Archaeoglobaceae; Archaeoglobus
Halobacterium sp. NRC-1, complete genome	NC_002607	2043086	cellular organisms; Archaea; Euryarchaeota; Halobacteria; Halobacteriales; Halobacteriaceae; Halobacterium
Methanococcus maripaludis C5, complete genome	NC_009135	1806279	cellular organisms; Archaea; Euryarchaeota; Methanococci; Methanococcales; Methanococcaceae; Methanococcus; Methanococcus maripaludis
Methanopyrus kandleri AV19, complete genome	NC_003551	1719258	cellular organisms; Archaea; Euryarchaeota; Methanopyri; Methanopyrales; Methanopyraceae; Methanopyrus; Methanopyrus kandleri

Thermoplasma volcanium GSS1, complete genome	NC_002689	1607521	cellular organisms; Archaea; Euryarchaeota; Thermoplasmata; Thermoplasmatales; Thermoplasmataceae; Thermoplasma; Thermoplasma volcanium
Methanospirillum hungatei JF-1, complete genome	NC_007796	3595457	cellular organisms; Archaea; Euryarchaeota; Methanomicrobia; Methanomicrobiales; Methanospirillaceae; Methanospirillum; Methanospirillum hungatei
Nanoarchaeum equitans Kin4-M, complete genome	NC_005213	497975	cellular organisms; Archaea; Nanoarchaeota; Nanoarchaeum; Nanoarchaeum equitans
Frankia sp. EAN1pec, complete genome	NC_009921	8982042	cellular organisms; Bacteria; Actinobacteria; Actinobacteria (class); Actinobacteridae; Actinomycetales; Frankineae; Frankiaceae; Frankia
Streptomyces avermitilis MA-4680, complete genome	NC_003155	9025608	cellular organisms; Bacteria; Actinobacteria; Actinobacteria (class); Actinobacteridae; Actinomycetales; Streptomycineae; Streptomycetaceae; Streptomyces
Aquifex aeolicus VF5, complete genome	NC_000918	1551335	cellular organisms; Bacteria; Aquificae; Aquificae (class); Aquificales; Aquificaceae; Aquifex
Acaryochloris marina MBIC11017, complete genome	NC_009925	6503724	cellular organisms; Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Sphingobacteria; Sphingobacteriales; Flexibacteraceae; Cytophaga; Cytophaga hutchinsonii
Chlamydophila pneumoniae CWL029, complete genome	NC_000922	1230230	cellular organisms; Bacteria; Chlamydiae/ Verrucomicrobia group; Chlamydiae; Chlamydiae (class); Chlamydiales; Chlamydiaceae; Chlamydophila; Chlamydophila pneumoniae
Herpetosiphon aurantiacus ATCC 23779, complete genome	NC_009972	6346587	cellular organisms; Bacteria; Chloroflexi; Chloroflexi (class); Herpetosiphonales; Herpetosiphonaceae; Herpetosiphon; Herpetosiphon aurantiacus
Nostoc sp. PCC 7120, complete genome	NC_003272	6413771	cellular organisms; Bacteria; Cyanobacteria; Nostocales; Nostocaceae
Deinococcus radiodurans R1 chromosome 1, complete sequence	NC_001263	2648638	cellular organisms; Bacteria; Deinococcus-Thermus; Deinococci; Deinococcales; Deinococcaceae; Deinococcus; Deinococcus radiodurans

Solibacter usitatus Ellin6076, complete genome	NC_008536	9965640	cellular organisms; Bacteria; Fibrobacteres /Acidobacteria group; Acidobacteria; Solibacteres; Solibacterales; Solibacteraceae; Solibacter
Alkaliphilus metalliredigens QYMF, complete genome	NC_009633	4929566	cellular organisms; Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridiaceae
Bacillus cereus ATCC 14579, complete genome	NC_004722	5411809	cellular organisms; Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus cereus group
Fusobacterium nucleatum subsp. nucleatum ATCC 25586, complete genome	NC_003454	2174500	cellular organisms; Bacteria; Fusobacteria; Fusobacteria (class); Fusobacteriales; Fusobacteriaceae; Fusobacterium; Fusobacterium nucleatum
Rhodopirellula baltica SH 1, complete genome	NC_005027	7145576	cellular organisms; Bacteria; Planctomycetes; Planctomycetacia; Planctomycetales; Planctomycetaceae; Rhodopirellula
Bradyrhizobium japonicum USDA 110, complete genome	NC_004463	9105828	cellular organisms; Bacteria; Proteobacteria; α-proteobacteria; Rhizobiales; Bradyrhizobiaceae; Bradyrhizobium; Bradyrhizobium japonicum
Delftia acidovorans SPH-1, complete genome	NC_010002	6767514	cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Delftia; Delftia acidovorans
Syntrophobacter fumaroxidans MPOB, complete genome	NC_008554	13033779	cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Myxococcales; Sorangineae; Polyangiaceae; Sorangium
Hahella chejuensis KCTC 2396, complete genome	NC_007645	7215267	cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Hahellaceae; Hahella; Hahella chejuensis
Leptospira interrogans serovar Lai str. 56601 chromosome I, complete sequence	NC_004342	4332241	cellular organisms; Bacteria; Spirochaetes; Spirochaetes (class); Spirochaetales; Leptospiraceae; Leptospira; Leptospira interrogans

Caldicellulosiruptor saccharolyticus DSM 8903, complete genome	NC_009437	2970275	cellular organisms; Bacteria; Synergistetes; Syntrophomonadaceae; Caldicellulosiruptor; Caldicellulosiruptor saccharolyticus
Petrotoga mobilis SJ95, complete genome	NC_010003	2169548	cellular organisms; Bacteria; Thermotogae; Thermotogae (class); Thermotogales; Thermotogaceae; Petrotoga; Petrotoga mobilis
Plasmodium falciparum 3D7 chromosome 14, complete sequence	NC_004317	3291006	cellular organisms; Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Haemosporida; Plasmodium; Plasmodium (Laverania)
Eimeria tenella str. Houghton chromosome 1, complete sequence	NC_008685	1347714	cellular organisms; Eukaryota; Alveolata; Apicomplexa; Coccidia; Eucoccidiorida; Eimeriorina; Eimeriidae; Eimeria
Paramecium tetraurelia macronuclear, complete genome	NC_006058	984602	cellular organisms; Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Oligohymenophorea; Peniculida; Parameciidae; Paramecium
Guillardia theta nucleomorph chromosome 1, complete sequence	NC_002752	196216	cellular organisms; Eukaryota; Cryptophyta; Cryptomonadaceae; Guillardia
Leishmaniabraziliensis MHOM/BR75/M2904 chromosome 20	NC_009312	1668259	cellular organisms; Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Leishmania; Viannia; Leishmania braziliensis species complex
Magnaporthe grisea 70-15 chromosome 7, complete sequence	NC_009594	3994966	cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Pezizomycotina; Sordariomycetes; Sordariomycetes incertae sedis; Magnaporthaceae; Magnaporthe; Magnaporthe grisea
Saccharomyces cerevisiae chromosome IV, complete chromosome sequence	NC_001136	1531918	cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces
Cryptococcus neoformans var. neoformans JEC21 chromosome 1, complete sequence	NC_006670	2300533	cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Basidiomycota; Agaricomycotina; Tremellomycetes; Tremellales; Tremellaceae; Filobasidiella

Encephalitozoon cuniculi GB-M1 chromosome XI, complete sequence	NC_003237	267509	cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Microsporidia; Apansporoblastina; Unikaryonidae; Encephalitozoon
Rattus norvegicus chromosome 12, reference assembly (based on RGSC v3.4)	NC_005111	46782294	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Rattus
Homosapiens chromosome 21, alternate assembly (based on Celera assembly), whole genome shotgun sequence	AC_000064	33216610	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homo/Pan/Gorilla group; Homo
Equus caballus chromosome 13, reference assembly (based on EquCab1), whole genome shotgun sequence	NC_009156	17519737	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Laurasiatheria; Perissodactyla; Equidae; Equus; Equus subg. Equus
Gallus gallus chromosome 9, reference assembly (based on Gallus gallus-2.1)	NC_006096	25554352	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Sauropsida; Sauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus

Danio rerio chromosome 25, reference assembly (based on Zv6)	NC_007136	40315040	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Elopocephala; Clupeocephala; Otocephala; Ostariophysi; Otophysi; Cypriniphysi; Cypriniformes; Cyprinoidea; Cyprinidae; Rasborinae; Danio
Tribolium castaneum linkage group 9, reference assembly (based on Tcas 2.0), whole genome shotgun sequence	NC_007424	15222296	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Protostomia; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Coleoptera; Polyphaga; Cucujiformia; Tenebrionoidea; Tenebrionidae; Tribolium
Caenorhabditis elegans chromosome III, complete sequence	NC_003281	13783681	cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Pseudocoelomata; Nematoda; Chromadorea; Rhabditida; Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis
Dictyostelium discoideum AX4 chromosome 2, complete sequence	NC_007088	8470428	cellular organisms; Eukaryota; Mycetozoa; Dictyosteliida; Dictyostelium
Ostreococcus lucimarinus CCE9901 chromosome 1, complete sequence	NC_009355	1152508	cellular organisms; Eukaryota; Viridiplantae; Chlorophyta; Prasinophyceae; Mamiellales; Mamiellaceae; Ostreococcus; Ostreococcus 'lucimarinus'
Arabidopsis thaliana chromosome 4, complete sequence	NC_003075	18585042	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; eurosids II; Brassicales; Brassicaceae; Arabidopsis

Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 10	NC_008403	22685906	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; Liliopsida; commelinids; Poales; Poaceae; BEP clade; Ehrhartoideae; Oryzeae; Oryza; Oryza sativa

The objective of retesting the accuracy of origin prediction with a larger diverse database is twofold. First, it facilitates the study of the behavior of all signatures when the database is larger. Second, a uniform sampling of organisms from different parts of the taxonomic tree will shed light on whether each signature is conserved for species in all parts of the tree and help identify pockets of high or low conservation of each signature in the tree. We computed databases of θ_2^{dbc}, θ_2^{dor}, θ_2^{wcv}, and θ_2^{combo} signatures of the 50 species in Table 6.3. Accuracies were determined for sequence samples at lengths 100 kb, 50 kb, 25 kb, 10kb, 5 kb, and 2.5 kb. As before, at each length, 100 subsequences of that length were randomly sampled from each genomic sequence listed in Table 6.3. For each sample, all signatures of order 2 were computed. The origin was predicted for each sample with each signature using its respective database. The accuracy at each length for each species was computed as the number of correct first hits in the 100 samples. Figure 6.15 illustrates the accuracy of the θ_2^{dbc} signature for each of the 50 species. The first 10 ticks on the x-axis correspond to archaeal genomic sequences, ticks 11-30 correspond to bacterial genomic sequences, and ticks 31-50 correspond to eukaryotic genomic sequences.

Observe that, as expected, accuracy drops with decreasing sequence length. The average accuracies of origin prediction of the θ_2^{dbc} signature for sequences of lengths 100 kb, 50 kb, 25 kb, 10 kb, 5 kb, and 2.5 kb are 95.3%, 94.04%, 92.52%, 87.02%, 80.52%, and 70.3%, respectively. We also observe that the accuracy of the θ_2^{dbc} signature is very low for the genomic sequences of *H. sapiens*, *E. caballus*, and *G. gallus*. The average accuracies of the θ_2^{dbc} signature for the above 3 species are 43.33%, 40.33%, 44.33%, 36%, 30.33%, and 25.33% for samples of lengths 100 kb, 50 kb, 25 kb, 10 kb, 5 kb, and 2.5 kb, respectively. Excluding these 3 species, the average accuracies of origin prediction of the θ_2^{dbc} signature for sequences of lengths 100 kb, 50 kb, 25 kb, 10 kb, 5 kb, and 2.5 kb are 98.62%, 97.47%, 95.60%, 90.28%, 83.68%, and 73.17%, respectively.

In Figures 6.16 and 6.17, we compare the accuracies of all the three signatures for all sample lengths listed above.

Observe that, the results are similar to those observed in Section 6.2.2. For samples of length 100 kb all three signatures demonstrate an average accuracies greater than 90%. As the sample length decreases,

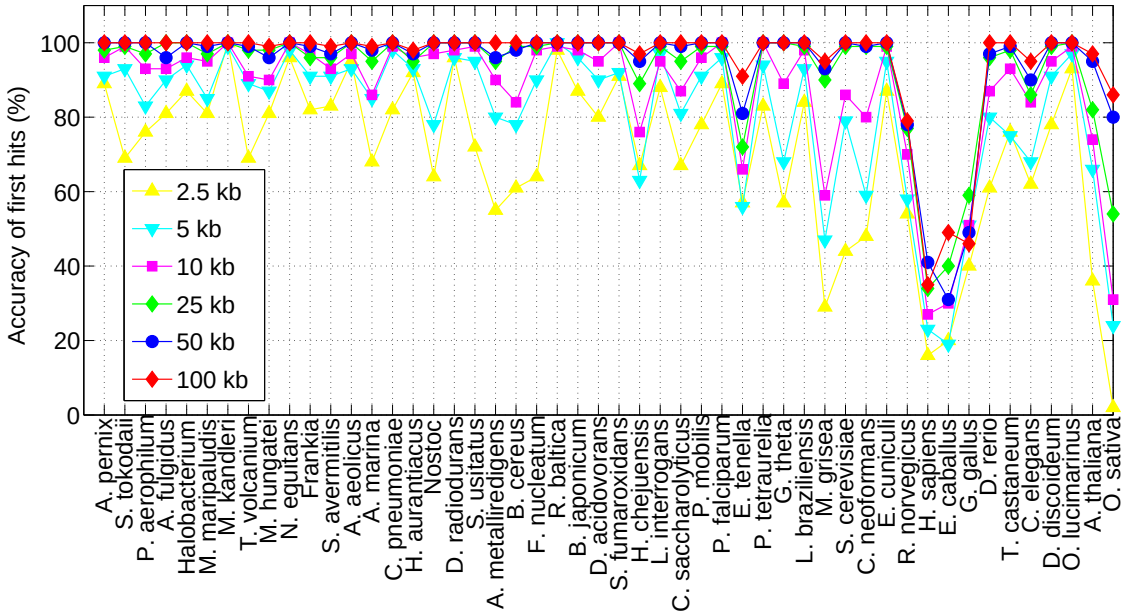
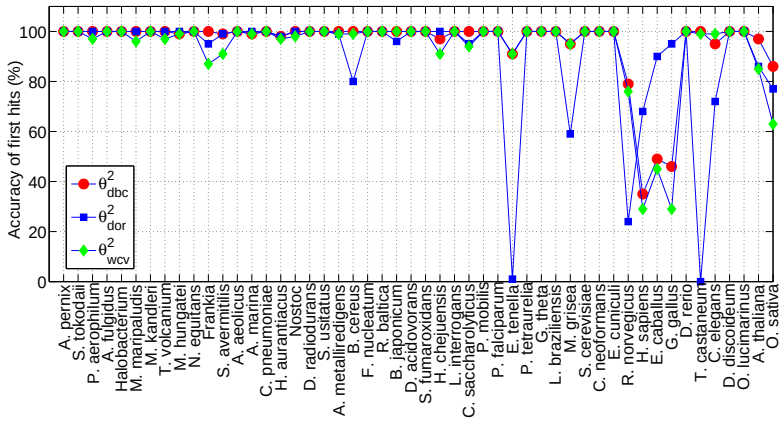


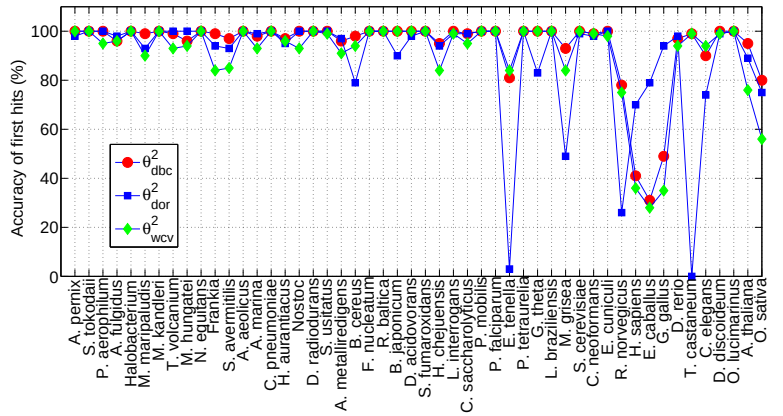
Figure 6.15: Accuracy of origin prediction of the θ_2^{dbc} signature for a large database. Accuracy of origin prediction of the θ_2^{dbc} signature using the species listed in Table 6.3.

accuracy falls for all signatures. We note that the θ_2^{dor} signature has a much higher fluctuation in accuracies among different species than the θ_2^{dbc} signature. For the 3 species mentioned in the previous paragraph, all three signatures demonstrate relatively lower accuracies. It is easy to see that in general, the θ_2^{dbc} signature demonstrates higher overall accuracy of prediction than the θ_2^{dor} and θ_2^{wcv} signatures. Specific numbers are discussed in the following passages. Also note that, for some species, the θ_2^{dbc} signature is consistently more well-conserved than the the θ_2^{dor} signature. These species are *A. fulgidus*, *M. maripaludis*, *T. volcanium*, *N. equitans*, *Frankia*, *S. avermitilis*, *A. marina*, *C. pneumoniae*, *H. aurantiacus*, *Nostoc*, *D. radiodurans*, *S. usitatus*, *A. metalliredigens*, *B. cereus*, *R. baltica*, *B. japonicum*, *D. acidovorans*, *S. fumaroxidans*, *L. interrogans*, *C. saccharolyticus*, *P. mobilis*, *P. falciparum*, *E. tenella*, *P. tetraurelia*, *G. theta*, *M. grisea*, *S. cerevisiae*, *R. norvegicus*, *T. castaneum*, *C. elegans*, *D. discoideum*, *O. lucimarinus*, *A. thaliana*, and *O. sativa*. For other species such as *A. pernix*, *P. aerophilum*, *M. hungatei*, *F. nucleatum*, *H. chejuensis*, *C. neoformans*, *E. cuniculi*, *H. sapiens*, *E. caballus*, *G. gallus*, and *D. rerio*, the θ_2^{dor} signature is consistently more well-conserved than the the θ_2^{dbc} signature.

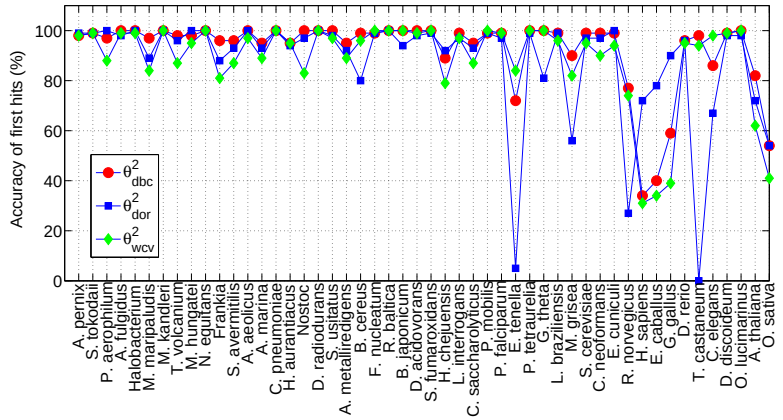
At sequence length of 5 kb, the accuracy of the θ_2^{dbc} signature is approximately 80%. Below 5 kb, the accuracy of the θ_2^{dbc} signature falls to values lower than 80%. The median accuracies of the three signatures for various sequence lengths are summarized in Figure 6.18. Observe that the median accuracy of the θ_2^{dbc}



(a)

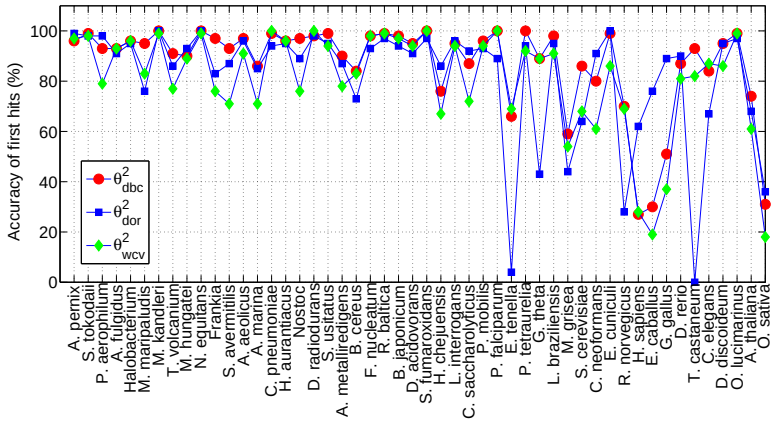


(b)

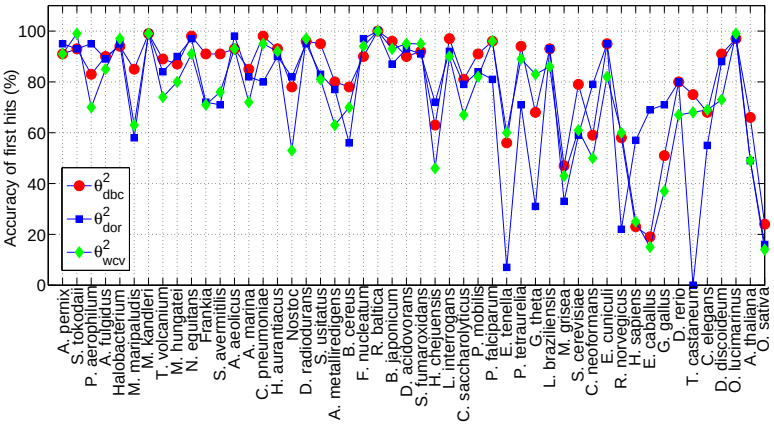


(c)

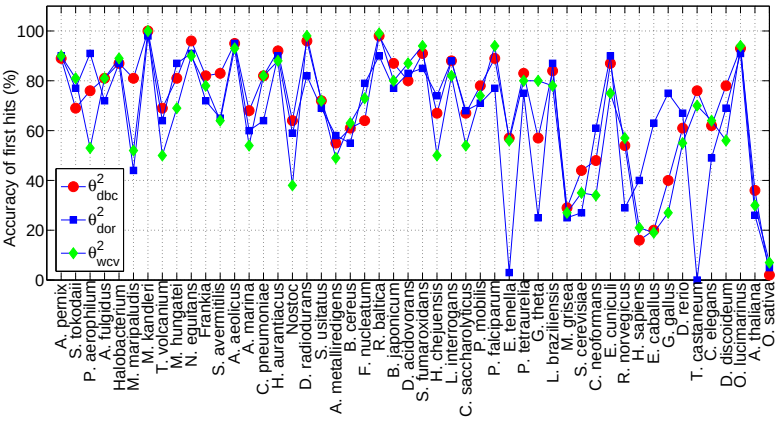
Figure 6.16: Accuracy of θ_2^{dbc}, θ_2^{dor}, and θ_2^{wcv} using a large database (i). Sample sequences of lengths (a) 100 kb, (b) 50 kb, and (c) 25 kb have been used.



(a)



(b)



(c)

Figure 6.17: Accuracy of θ_2^{dbc}, θ_2^{dor}, and θ_2^{wcv} using a large database (ii). Sample sequences of lengths (a) 10 kb, (b) 5 kb, and (c) 2.5 kb have been used.

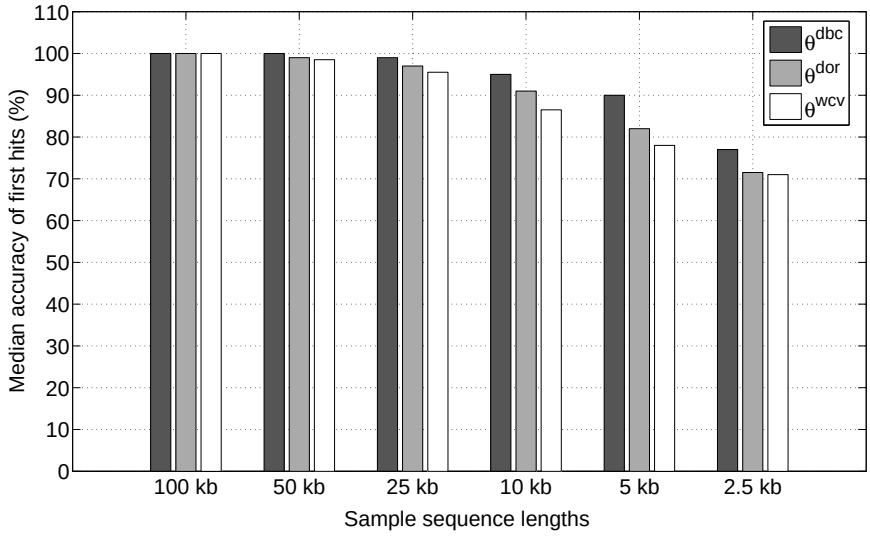


Figure 6.18: Median accuracies of θ_2^{dbc}, θ_2^{dor}, and θ_2^{wcv} using a large database.

signature is higher than those of the other signatures at every sample length. Also observe that, as the sample length decreases, the amount by which the median accuracy of the θ_2^{dbc} signature is greater than those of the other signatures increases.

The experiments conducted in this section demonstrate that the θ_2^{dbc} signature predicts origins of short DNA segments accurately even when the database size is large, and in doing so, performs much better than θ_2^{wcv} and θ_2^{dor} genomic signatures.

6.2.5 Relationship between genome size and accuracy of origin prediction

Next, we explored pairwise relationships between genome size, genome variation, and accuracy of first hits of the θ_2^{dbc} signature using sequence samples of length 10 Kb. Given a genomic sequence H, define the *genome variation* of H as follows. Define an order w, a window length W, and a skip length s. Compute the word frequency vector signature $\theta_w^{wfv}(H)$ for the entire genome. Start at the beginning of the sequence and read the sequence W characters at a time, while sliding the window by s characters each time. For each substring thus read, compute the word frequency vector signature and store the component-wise absolute difference from $\theta_w^{wfv}(H)$. After the sliding window has read the entire sequence, compute the average absolute difference for each component. The *genome variation* is the sum of the averages thus computed. Figure 6.19 presents scatter plots for 3 kinds of pairwise relationships possible. We have used $W = 10$ kb and $s = 2.5$ kb for our

computations. Plots (a), (b), and (c) are for 11 out of 12 genomes in list L_1 (the human genome was not used as it was an outlier that disrupted the otherwise observed correlations, because of its large size), while plots (d), (e), and (f) are for the 20 α-proteobacterial species. Observe that the accuracy of θ^{dbc} is negatively correlated with genome size in both sets, using the Pearson correlation coefficient. Other relationships are not obvious from these results.

However, when we study the variation of accuracy with genome size for the larger database of 50 species, the negative correlation is not observed. The results are shown in Figure 6.20.

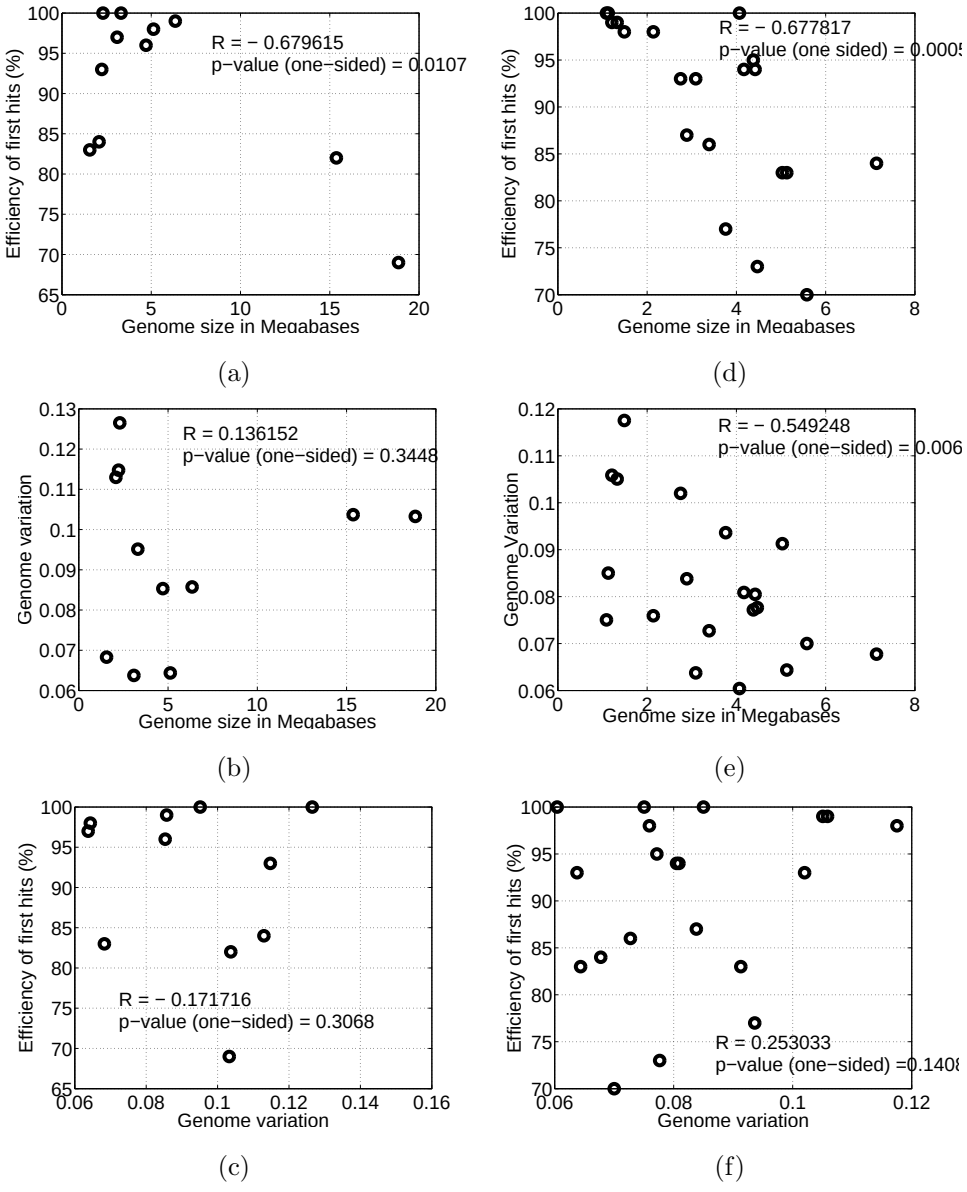


Figure 6.19: Relationships between genome size, genome variation, and accuracy of θ_2^{dbc}. (a), (b), and (c) demonstrate results for the first 11 organisms in list L_1. (d), (e), and (f) demonstrate results for the 20 α-proteobacterial species.

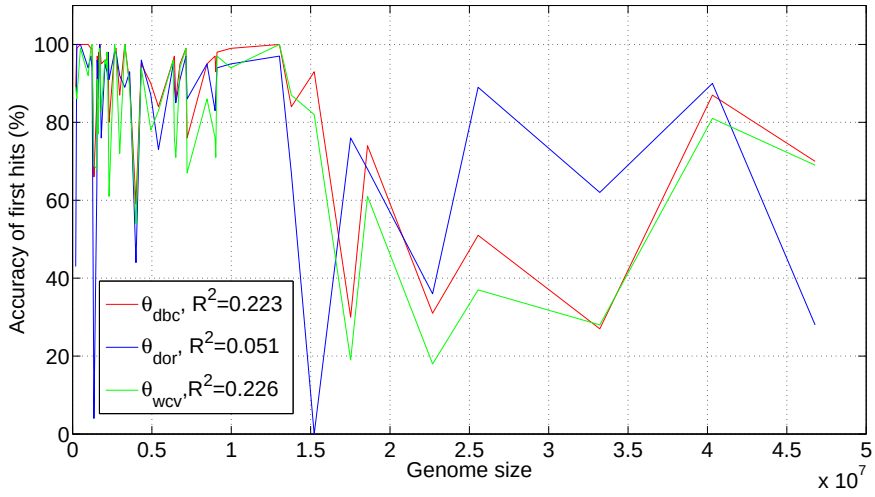


Figure 6.20: Variation of accuracy with genome size for 50 species. All three signatures have been examined.

Chapter 7

Estimating Markov Chain Order

7.1 Introduction

At its lowest level, almost every organizational unit of a genome is a genomic (that is, contiguous) sequence, which is formally a string over the alphabet $\Sigma_{\text{DNA}} = \{A, C, G, T\}$. Genomic sequences encode the genes of an organism, among other characteristics. We hypothesize that each segment of a genome is generated, within a reasonable approximation, by a Markov chain of unknown order. Given a sequence S, we call such a Markov chain $\mathcal{M}$ that generates S, the *generating Markov chain* of S. Estimating the order of the generating Markov chain will assist in understanding biological phenomena such as a difference in frequencies of observed DNA word[1] patterns and repeats in the genome [52, 54].

Others [3, 25, 28, 36, 95, 110] have proposed methods for estimating the order of the generating Markov chain of a sequence. Peres and Shields [95] introduce two Markov order estimators. Both estimators use test functions that depend on sample size and a candidate w for the order. As w increases, the test functions exhibit a qualitative change of behavior when w reaches the true order. The first test function is based on a form of entropy, while the second test function is based on maximal fluctuation (see Section 7.2.3) and is more relevant to our formulation.

Dalevi and Dubhashi [28] give a novel interpretation of the Peres-Shields estimator as a sharp transition function. They claim that their interpretation makes the estimator more useful in the context of DNA sequences when sequence sizes are moderate, and extend the estimator to variable length Markov chains.

[1] A *DNA word* of length w is a string in Σ_{DNA}^w.

Their method is useful in identifying the order of the generating Markov chain for a sequence effectively. However, a mathematical framework that models nucleotide frequencies and the sharp transition function was not proposed. Moreover, the algorithm has time complexity that is exponential in the order of the underlying Markov chain.

Existing methods are based on principles of entropy estimation, maximal fluctuation, and maximum likelihood estimation. In this work, we propose a randomized algorithm for estimating the order of a generating Markov chain within a framework of probability distributions of its states and transitions.

Section 7.2 introduces and defines relevant computational concepts, and establishes notation. Section 7.3 builds the framework, computes a distribution for the transition probability, and describes the algorithm to estimate order. We give a qualitative description of the maximum fluctuation that leads to a decision about the generating Markov chain order. Section 7.5 discusses results and illustrates the maximum fluctuation. In Section 7.6, we draw conclusions and discuss possible improvements and future directions.

7.2 Preliminaries

7.2.1 Strings

As usual, an *alphabet* Σ is a finite, nonempty set of symbols. A *string* over Σ is a finite sequence of symbols from Σ. Henceforth, we take $\Sigma = \Sigma_{\text{DNA}} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$. The *length* of a string is its length as a sequence. The *empty string* λ is the unique string of length 0. For $n \geq 0$, Σ^n is the set of strings of length n over Σ, while $\Sigma^* = \bigcup_{n \geq 0} \Sigma^n$ is the set of all strings over Σ. By convention, we employ α, β, and γ for strings and ρ, σ, and τ for symbols.

The concatenation of strings α and β is $\alpha \cdot \beta$ or, simply, $\alpha\beta$. Let $\alpha = \rho_1\rho_2 \cdots \rho_n \in \Sigma^n$. For $1 \leq k \leq n$, the k^{th} *character* of α is $\alpha[k] = \rho_k$. If $1 \leq i \leq j \leq n$, then the (i, j) *substring or subsequence* of α is $\alpha[i..j] = \rho_i\rho_{i+1} \cdots \rho_j$; otherwise, $\alpha[i..j] = \lambda$. The (i, j) substring $\alpha[i..j]$ *occurs at position* i. For strings α and β, the predicate $\beta \stackrel{i}{\sqsupset} \alpha$ is true just when $\beta = \alpha[i..j]$, for some $j \geq 0$.

Let $\alpha = \rho_1\rho_2 \cdots \rho_n \in \Sigma^n$. For $0 \leq k \leq n$, the *length-k prefix* of α is $\alpha[1..k] = \rho_1\rho_2 \cdots \rho_k$, while the *length-k suffix* of α is $\alpha[n - k + 1..n] = \rho_{n-k+1}\rho_{n-k+2} \cdots \rho_n$. Strings α and β *overlap* if a nonempty prefix of α is a suffix of β or vice versa. Strings α and β are *non-overlapping* if they do not overlap.

The *count of the occurrences of* β *in* α is

$$\Psi(\alpha, \beta) \quad = \quad \left| \left\{ i \mid 1 \leq i \leq n - |\beta| + 1 \text{ and } \beta \stackrel{i}{\sqsupset} \alpha \right\} \right|.$$

The *frequency* of β in α is

$$\text{freq}\,(\alpha,\beta) \quad = \quad \begin{cases} \dfrac{\Psi(\alpha,\beta)}{n-|\beta|+1} & \text{if } n \geq |\beta|; \\ 0 & \text{otherwise.} \end{cases}$$

7.2.2 Probabilities

For $w \geq 1$, a *Markov chain* $\mathcal{M} = (\Sigma^w, P)$ *of order* w *over* Σ consists of the *state space* Σ^w and a $4^w \times 4^w$ stochastic matrix P of *transition probabilities* with rows and columns indexed by elements of Σ^w. For $\alpha, \beta \in \Sigma^w$ satisfying $\alpha[2,w] = \beta[1..w-1]$, the transition probability $p_{\alpha,\beta}$ is the *probability of leaving* α *on the symbol* $\beta[w]$. Alternately, if $\sigma \in \Sigma$, then the transition from $\alpha \in \Sigma^w$ to $\beta = \alpha[2..w] \cdot \sigma$ is abbreviated $\alpha \xrightarrow{\sigma} \beta$ and $p_{\alpha,\beta} = \Pr\left[\alpha \xrightarrow{\sigma} \beta\right]$.

Let $\mathcal{M}$ be an ergodic Markov chain of order w. Let S be a random sequence generated by $\mathcal{M}$. The Markov Order problem is to determine the order of $\mathcal{M}$, using only the generated sequence S. For $\alpha \xrightarrow{\sigma} \beta$, the *empirical transition probability* from α to β is

$$p_{\text{e}}\,(\alpha,\beta) = \begin{cases} \frac{\Psi(S,\alpha\cdot\sigma)}{\Psi(S,\alpha)} & \text{if } \Psi(S,\alpha) \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

As an example, let $w = 3$, $\alpha = \text{AAG}$, $\beta = \text{AGT}$, and

$$S = \text{AAGTCGAAGTTATGTCGGTAAGCCAGCGCCCAAGA.}$$

By observation, $\Psi(S, \text{AAG}) = 4$ and $\Psi(S, \text{AAGT}) = 2$. Hence, $p_{\text{e}}\,(\text{AAG}, \text{AGT}) = 2/4 = 1/2$. The *derived transition probability* from α to β is the empirical transition probability for $\alpha[2,w] \xrightarrow{\sigma} \beta[2,w]$, which is

$$p_{\text{d}}\,(\alpha,\beta) \quad = \quad p_{\text{e}}\,(\alpha[2,w], \beta[2,w])\,.$$

In the previous example,

$$p_{\text{d}}\,(\text{AAG}, \text{AGT}) = p_{\text{e}}\,(\text{AG}, \text{GT}) = 2/5.$$

Clearly, $p_{\text{e}}\,(\alpha,\beta)$ and $p_{\text{d}}\,(\alpha,\beta)$ may differ.

Note that both the empirical and derived transition probabilities give transition matrices for Markov chains of order w. The *probability differential* Δ_w is the L1 distance between the two transition matrices:

$$\Delta_w \quad = \quad \sum_{\alpha\in\Sigma^w} \sum_{\sigma\in\Sigma} \sum_{\alpha\xrightarrow{\sigma}\beta} |p_{\text{e}}\,(\alpha,\beta) - p_{\text{d}}\,(\alpha,\beta)|\,.$$

Intuitively, for any β randomly chosen, this is very small if $w-1$ is the order of the Markov chain that generated S. To empirically confirm this intuition, sequences of length between 50 and 19531250 were

Table 7.1: Distances between empirical and derived probability distributions: true order 5. Distances between empirical and derived probability distributions of a sequence of length 781250 bases generated by a Markov chain of order 5.

w	d_{BC}	d_{KL}	d_{L1}	d_{L2}	d_{cos}
3	0.01215	0.33447	0.19453	0.06930	0.03565
4	0.00529	0.45641	0.16951	0.04048	0.02314
5	0.00496	3.53444	0.31770	0.05248	0.07102
6	3.55E-05	0.00249	0.00454	0.00063	2.03E-05
7	5.75E-05	0.13273	0.01472	0.00258	0.00066
8	4.04E-05	0.32663	0.02053	0.00191	0.00072
9	3.30E-05	5.00267	0.03317	0.00281	0.00314
10	2.59E-05	15.81827	0.05098	0.00264	0.00560

generated by Markov chains of various orders. These sequences were then used to estimate the order of the generating Markov chain using distances between empirical and derived distributions at each order. Table 7.1 summarizes the results for a sequence of length 781250 bases that was generated using a Markov chain of order 5. Table 7.2 summarizes the results for a sequence of length 3906250 bases that was generated using a Markov chain of order 4. Observe that when the value of w equals one more than the order of the generating Markov chain, i.e., $w = 6$ in Table 7.1 and $w = 5$ in Table 7.2, every distance between the empirical and derived distributions falls abruptly to a very small number.

7.2.3 Maximal Fluctuations

Let $S = \sigma_1\sigma_2\cdots$ be an infinite sequence generated by a Markov chain of order $\hat{w}$. For $n \geq 1$, let $f_n : \Sigma^n \to \mathbb{N}$ be any function. Peres and Shields [95] define the sequence $f_1, f_2, \ldots$ to be a *consistent order estimator* if $\lim f_n(S[1..n]) = w$ with probability 1.

Dalevi and Dubhashi [28] interpret the Peres-Shields estimator as a sharp transition function described below. Let $\beta \in \Sigma^l$ and $\gamma = \beta[l - w + 1 \ldots l]$ be the w-suffix of β. Then, given a sequence S and for $\sigma \in \Sigma,$, $D_S^w(\beta)$ is defined as the difference between the observed and expected number of occurrences of $\beta\sigma$ in S as follows:

$$D_S^w(\beta) \quad = \quad \max_{\sigma\in\Sigma} \left| \Psi(S, \beta\sigma) - \frac{\Psi(S, \gamma\sigma)}{\Psi(S, \gamma)} \Psi(S, \beta) \right|,$$

Table 7.2: Distances between empirical and derived probability distributions: true order 4. Distances between empirical and derived probability distributions of a sequence of length 3906250 bases generated by a Markov chain of order 4.

w	d_{BC}	d_{KL}	d_{L1}	d_{L2}	d_{cos}
3	0.02765	1.27104	0.44251	0.13007	0.10903
4	0.00756	1.21845	0.24209	0.05821	0.04057
5	2.56545e-05	0.00013	0.00164	0.00030	2.12975e-06
6	4.37504e-05	0.00633	0.00560	0.00105	5.20785e-05
7	4.04292e-05	0.05512	0.01034	0.00155	0.00022
8	3.94200e-05	1.25930	0.02014	0.00254	0.00120
9	3.15086e-05	3.62831	0.03190	0.00279	0.00293
10	2.15628e-05	14.67987	0.04258	0.00251	0.00481

The estimator is described as

$$w_{PS}(S) \quad = \quad \min\left\{ w \geq 0 | \max_{w<|\beta|<\log\log n} D_S^w(\beta) < n^{3/4} \right\},$$

where they state that this interpretation makes the estimator more useful in the context of DNA sequences when sequence sizes are moderate. They also extend the estimator to variable length Markov chains.

7.3 Theory

Let $\mathcal{M} = (\Sigma^{\hat{w}}, P)$ be an ergodic Markov chain of order $\hat{w}$. Let S be a random sequence of length $n > \hat{w}$ generated by $\mathcal{M}$. For some w satisfying $1 \leq w \leq n-1$, let β be a random string of length $w+1$ that occurs in S and does not overlap itself. Let $\alpha = \beta[1..w]$, the length w prefix of β, and let $\gamma = \beta[2..w+1]$, the length w suffix of β. Let $\sigma = \beta[w+1]$. Then $\alpha \xrightarrow{\sigma} \gamma$.

For $1 \leq i \leq n$, let X_i be the indicator random variable for $\beta \overset{i}{\sqsupset} S$ and Y_i the indicator random variable for $\alpha \overset{i}{\sqsupset} S$. Then, $X = \sum_i X_i = \Psi(S, \beta)$ and $Y = \sum_i Y_i = \Psi(S, \alpha)$ are non-negative random variables. Note that $X_i = 1$ implies $Y_i = 1$, for all i; therefore, $0 \leq X \leq Y$.

We assume that $\Pr[X_i = 1 \mid Y_i = 1] = p_{\mathrm{e}}(\alpha, \gamma) = X/Y$. For purposes of abbreviation, set $p = p_{\mathrm{e}}(\alpha, \gamma)$. Let $f_Y : \mathbb{N} \to \mathbb{R}$ be the probability density function for Y. Note that, since β occurs in S, we have $f_Y(0) = 0$.

Define Z to be the random variable

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

Then Lemma 7.1 and Lemma 7.2 hold.

Lemma 7.1. *Under the assumption that X_i conditioned on Y_i occurs with probability p, we have*

$$\mathbf{E}[X] = p\mathbf{E}[Y]$$

and

$$\mathrm{Var}[X] = p\mathrm{Var}[Y] + p(1-p)(\mathbf{E}[Y])^2.$$

Proof. Start with the probability distribution of X, which is

$$\begin{aligned} f_X(x) &= \Pr[X = x] \\ &= \sum_{y \geq x} \Pr[X = x \mid Y = y] \Pr[Y = y]. \end{aligned}$$

The expected value is now

$$\begin{aligned} \mathbf{E}[X] &= \sum_{x \geq 0} x \sum_{y \geq x} \Pr[X = x \mid Y = y] \Pr[Y = y] \\ &= \sum_{x \geq 0} x \sum_{y \geq x} \binom{y}{x} p^x (1-p)^{y-x} \Pr[Y = y] \\ &= \sum_{y \geq 0} \Pr[Y = y] \sum_{0 \leq x \leq y} x \binom{y}{x} p^x (1-p)^{y-x} \\ &= \sum_{y \geq 0} \Pr[Y = y] (yp) \\ &= p\mathbf{E}[Y]. \end{aligned}$$

The variance is then

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \\
&= \mathbf{E}\left[X^2\right] - p^2(\mathbf{E}[Y])^2 \\
&= \sum_{x\geq 0} x^2 \sum_{y\geq x} \Pr[X = x \mid Y = y] \Pr[Y = y] - p^2(\mathbf{E}[Y])^2 \\
&= \sum_{x\geq 0} x^2 \sum_{y\geq x} \binom{y}{x} p^x (1-p)^{y-x} \Pr[Y = y] - p^2(\mathbf{E}[Y])^2 \\
&= \sum_{y\geq 0} \Pr[Y = y] \sum_{0\leq x\leq y} x^2 \binom{y}{x} p^x (1-p)^{y-x} - p^2(\mathbf{E}[Y])^2 \\
&= \sum_{y\geq 0} \Pr[Y = y] \left(y^2 p(1-p) + (yp)^2\right) - p^2(\mathbf{E}[Y])^2 \\
&= p\mathbf{E}\left[Y^2\right] - p^2(\mathbf{E}[Y])^2 \\
&= p(\mathrm{Var}[Y] + (\mathbf{E}[Y])^2) - p^2(\mathbf{E}[Y])^2 \\
&= p\mathrm{Var}[Y] + p(1-p)(\mathbf{E}[Y])^2.
\end{aligned}$$

□

Lemma 7.2. *Under the assumption that X_i conditioned on Y_i occurs with probability p, we have*

$$\mathbf{E}[Z] = p$$

and

$$\mathrm{Var}[Z] = p(1-p)\mathbf{E}[1/Y].$$

Proof. Start with the probability distribution of Z, which is, for $0 < z \leq 1$ and z rational,

$$\begin{aligned}
f_Z(z) &= \Pr[Z = z] \\
&= \sum_{y\geq 1} \sum_{0\leq x\leq y} \Pr[X = yz \mid Y = y] \Pr[Y = y].
\end{aligned}$$

Note that, if yz is not an integer, then

$$\Pr[x = yz \mid Y = y] = 0.$$

The expected value is now

$$\begin{aligned}
\mathbf{E}[Z] &= \sum_{y\geq 1}\sum_{0\leq x\leq y}\left(\frac{x}{y}\right)\Pr[X=x\mid Y=y]\Pr[Y=y] \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y}\sum_{0\leq x\leq y} x\Pr[X=x\mid Y=y] \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y}\sum_{0\leq x\leq y} x\binom{y}{x}p^x(1-p)^{y-x} \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y}(yp) \\
&= p.
\end{aligned}$$

The variance is then

$$\begin{aligned}
\mathrm{Var}[Z] &= \mathbf{E}\left[Z^2\right]-(\mathbf{E}[Z])^2 \\
&= \mathbf{E}\left[Z^2\right]-p^2 \\
&= \sum_{y\geq 1}\sum_{0\leq x\leq y}\left(\frac{x^2}{y^2}\right)\Pr[X=x\mid Y=y]\Pr[Y=y]-p^2 \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y^2}\sum_{0\leq x\leq y} x^2\Pr[X=x\mid Y=y]-p^2 \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y^2}\sum_{0\leq x\leq y} x^2\binom{y}{x}p^x(1-p)^{y-x}-p^2 \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y^2}(yp(1-p)+y^2p^2)-p^2 \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y^2}(yp(1-p)+y^2p^2)-p^2 \\
&= \sum_{y\geq 1}\frac{\Pr[Y=y]}{y}(p(1-p))+\sum_{y\geq 1}\Pr[Y=y](p^2)-p^2 \\
&= p(1-p)\mathbf{E}[1/Y]+p^2-p^2 \\
&= p(1-p)\mathbf{E}[1/Y].
\end{aligned}$$

□

The proofs are straight-forward and have been derived from first principles. Next, we derive upper bounds on $\mathbf{E}[1/Y]$ as given in Lemmas 7.3 and 7.4.

Lemma 7.3. *If Y has a Poisson distribution with parameter λ, then*

$$\mathbf{E}[1/Y] < 2\lambda^{-1}-2e^{-\lambda}\lambda^{-1}-2e^{-\lambda}.$$

Proof. The upper bound is obtained as follows.

$$\begin{aligned}
\mathbf{E}\left[1/Y\right] &= e^{-\lambda}\sum_{y\geq 1}\frac{1}{y}\cdot\frac{\lambda^y}{y!} \\
&= e^{-\lambda}\lambda^{-1}\sum_{y\geq 1}\frac{y+1}{y}\cdot\frac{\lambda^{y+1}}{(y+1)!} \\
&= e^{-\lambda}\lambda^{-1}\sum_{y\geq 1}\left(1+\frac{1}{y}\right)\cdot\frac{\lambda^{y+1}}{(y+1)!} \\
&< e^{-\lambda}\lambda^{-1}\sum_{y\geq 1}2\cdot\frac{\lambda^{y+1}}{(y+1)!} \\
&= 2e^{-\lambda}\lambda^{-1}\left(e^{\lambda}-1-\lambda\right) \\
&= 2\lambda^{-1}-2e^{-\lambda}\lambda^{-1}-2e^{-\lambda}.
\end{aligned}$$

□

Lemma 7.4. *If Y has a binomial distribution with parameters m and p, then*

$$\mathbf{E}\left[1/Y\right] < \frac{2}{(m+1)p}\left(1-(1-p)^{m+1}-(m+1)p(1-p)^m\right).$$

Proof. The upper bound is obtained as follows.

$$\begin{aligned}
\mathbf{E}\left[1/Y\right] &= \sum_{y=1}^{m}\frac{1}{y}\binom{m}{y}p^y(1-p)^{m-y} \\
&= \sum_{y=1}^{m}\frac{1}{y}\frac{m!}{y!(m-y)!}\cdot p^y(1-p)^{m-y} \\
&= \sum_{y=1}^{m}\frac{y+1}{y}\frac{m!}{(y+1)!(m-y)!}\cdot p^y(1-p)^{m-y} \\
&< 2\sum_{y=1}^{m}\frac{m!}{(y+1)!(m-y)!}p^y(1-p)^{m-y} \\
&= \frac{2}{m+1}\sum_{y=2}^{m+1}\frac{(m+1)!}{y!(m+1-y)!}p^{y-1}(1-p)^{m+1-y} \\
&= \frac{2}{(m+1)p}\sum_{y=2}^{m+1}\frac{(m+1)!}{y!(m+1-y)!}p^y(1-p)^{m+1-y} \\
&= \frac{2}{(m+1)p}\left(1-(1-p)^{m+1}-(m+1)p(1-p)^m\right).
\end{aligned}$$

□

Let S be a random sequence of length n generated by $\mathcal{M}=(\Sigma^{\hat{w}},P)$. Let β be a random string of length $l>\hat{w}$ that occurs in S. Let $\sigma=\beta[l]$. For w satisfying $l-1\geq w>0$, let $\alpha_w=\beta[l-w\ldots l-1]$, the length

w suffix of the length $l-1$ prefix of β, and let $\gamma_w = \beta[l-w+1 \ldots l]$, the length w suffix of β. Similarly, let $\alpha_{w-1} = \beta[l-w+1 \ldots l-1]$, and $\gamma_{w-1} = \beta[l-w+2 \ldots l]$. Let Z_w be the random variable for the transition probability $p = p_{\mathrm{e}}(\alpha_w, \gamma_w)$ and let Z_{w-1} be the random variable for the transition probability $p_{\mathrm{e}}(\alpha_{w-1}, \gamma_{w-1})$. Let

$$\Delta_w \quad = \quad |Z_w - Z_{w-1}|.$$

Then, Theorem 7.5 follows.

Theorem 7.5. *Let S be a sequence of length n generated by a Markov chain $\mathcal{M}$ of order $\hat{w}$. If Δ_w is as defined above, then for all $w > \hat{w}$,*

$$\mathbf{E}[\Delta_w] \quad = \quad 0.$$

Proof. Since the order of $\mathcal{M}$ is $\hat{w}$, and $w > \hat{w}$, we have $w-1 \geq \hat{w}$. By definition of w and $w-1$, $\mathbf{E}[Z_w] = p$ and $\mathbf{E}[Z_{w-1}] = p$. Therefore,

$$\mathbf{E}[\Delta_w] = \mathbf{E}[Z_w - Z_{w-1}] = \mathbf{E}[Z_w] - \mathbf{E}[Z_{w-1}] = p - p = 0.$$

□

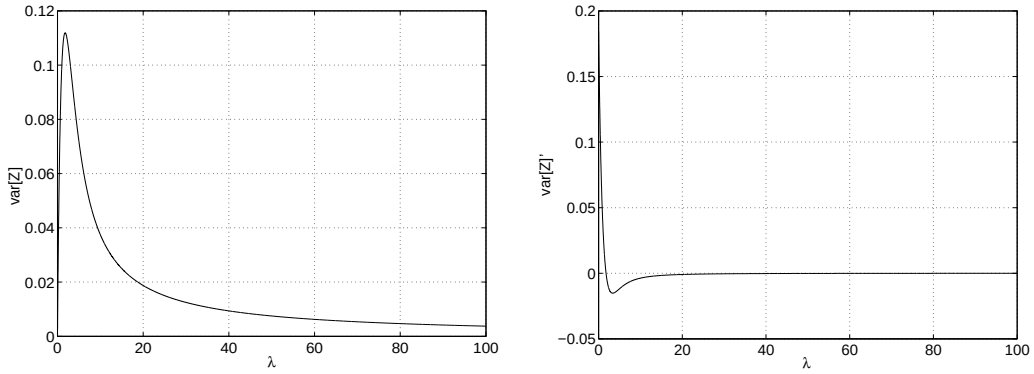
Consider $\mathrm{Var}[Z]$. If Y is Poisson distributed with parameter λ, and X_i, conditioned on Y_i, occurs with probability p, from Lemmas 7.2 and 7.3, $\mathrm{Var}[Z]$ has the following upper bound $\mathrm{Var_u}[Z]$:

$$\mathrm{Var_u}[Z] \quad = \quad 2p(1-p)(\lambda^{-1} - e^{-\lambda}\lambda^{-1} - e^{-\lambda}).$$

Figure 7.1 depicts the behavior of $\mathrm{Var_u}[Z]$ and $\mathrm{Var_u}[Z]'$ with increasing λ. $\mathrm{Var_u}[Z]' = \dfrac{d\mathrm{Var_u}[Z]}{d\lambda}$ is the function

$$2p(1-p)(e^{-\lambda} - \lambda^{-2} + \lambda^{-2}e^{-\lambda} + \lambda^{-1}e^{-\lambda}),$$

which has zeroes at $\lambda = \{0, 1.79328213\}$. Moreover, $\mathrm{Var_u}[Z]'$ decreases in the interval $\lambda \in (0, 1.79328213]$ and increases in the interval $\lambda \in [1.79328213, \infty)$. For a given α, let π_α be its stationary probability. Then, $\lambda_\alpha = n\pi_\alpha$. Observe that as the length of α increases, π_α decreases. In case of uniformly distributed words, π_α decreases approximately by a factor of 4 with each successive increase in the length of α by 1. For a sequence of fixed length, in the interval $[1.793, \infty)$, $\mathrm{Var_u}[Z]$ increases with decreasing λ, and hence π, while increasing with increasing word size. This is intuitive and expected. However, in the interval $[0, 1.793]$, $\mathrm{Var_u}[Z]$ increases with increasing λ, and hence π, and thus decreases with increasing word size. Although



(a) (b)

Figure 7.1: Behavior of $\mathrm{Var_u}[Z]$ and (b) $\mathrm{Var_u}[Z]'$ with λ and Poisson distributed Y. $\lambda = 1.79328213$ is a local maximum. $\mathrm{Var_u}[Z]$ decreases with increasing λ, therefore, increasing with increasing word length in the interval $[1.79328213, \infty)$, where it occurs with high probability. It displays opposite behavior in the interval $[0, 1.79328213]$, where its probability of occurrence is low.

the behavior of $\mathrm{Var_u}[Z]$ is inconsistent with what is expected, the probability of λ being in this interval is extremely low as shown in Lemma 7.6.

If Y is binomially distributed with parameters m and p', then $\mathrm{Var}[Z]$ has the following upper bound:

$$\mathrm{Var_u}[Z] \quad \leq \quad p(1-p)\frac{2}{(m+1)p'}(1-(1-p')^{m+1}-(m+1)p'(1-p')^m).$$

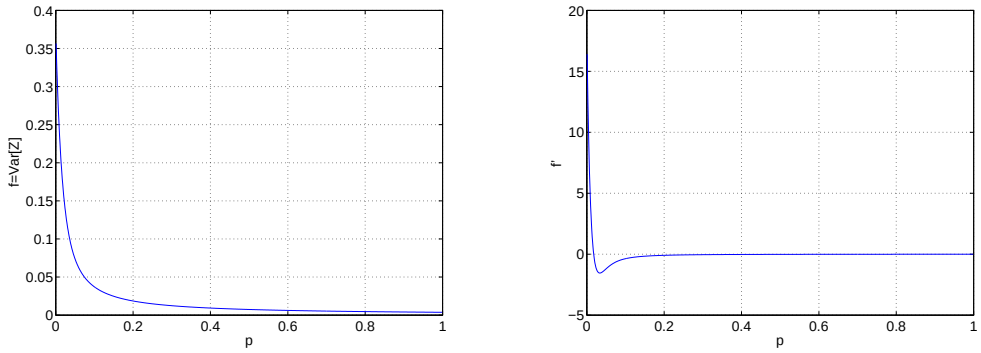
Figure 7.2 depicts the behavior of $\mathrm{Var_u}[Z]$ and $\mathrm{Var_u}[Z]'$ with increasing p'. Observe that the behavior exhibited by $\mathrm{Var_u}[Z]$ is similar in case of both the Poisson and binomial approximations to the distribution of Y.

Lemma 7.6. *Let S be a sequence of length $n = 4^k$. Assuming that $k >> 2w$, and S is observed at scale w, the probability that $\lambda \in [0, 2]$ is bounded above by*

$$\Pr[\lambda \in [0,2]] \quad \leq \quad \exp\left(-2 \cdot 4^{k-4w} - 4^{1-w} + 4^{1-k}\right).$$

Proof. There are 4^w words at scale w, and approximately 4^k (precisely $4^k - w + 1$) positions in S where a word could occur. The probability that any word has less than or equal to 2 occurrences is then given by the left tail of the binomial distribution as follows:

$$\Pr[\lambda \leq 2] \quad = \quad B\left(0; 4^k, 4^{-w}\right) + B\left(1; 4^k, 4^{-w}\right) + B\left(2; 4^k, 4^{-w}\right),$$



(a) (b)

Figure 7.2: Behavior of $\mathrm{Var_u}[Z]$ and (b) $\mathrm{Var_u}[Z]'$ with p' and binomially-distributed Y.

where $B(k; n, p)$ is the binomial probability of observing k successes in a Binomial distribution with parameters $(n, 4^{-w})$, and $B(k; n, p)$ is the corresponding cumulative probability. Since $k >> 2w$, applying Chernoff's inequality [20] to the binomial distribution, we have

$$\begin{aligned} \Pr[\lambda \leq 2] &= F\left(2; 4^k, \frac{1}{4^w}\right) \\ &\leq \exp\left(-2(4^{k-2w} - 4^{1-w} + 4^{1-k})\right). \end{aligned}$$

□

Figure 7.3 illustrates the probability variation for different values of k and w. It is clear that for $k >> 2w$, the probability is close to zero. The above results suggest that, for a sequence S of length n, that is analyzed at scale w, $\mathrm{Var_u}[Z]$ increases with increasing w with high probability.

Theorem 7.5 suggests that, for a sequence S generated by a Markov chain of order w, $|Z_k - Z_{k+1}|$ is very close to zero for all $k \geq w$. We use this fact to estimate the order of the generating Markov chain for a sequence as described in the algorithm in Figure 7.4.

The upper bound on the variance of Z, $\mathrm{Var_u}[Z]$ increases with increasing k for a given sequence. For a given $w_{\max}$ and a sequence S of length n, our algorithm estimates $\hat{w}$ in $O(\mu(nw_{\max} + c))$ time and space, where μ is the number of times β is sampled. Values of μ are discussed in Section 7.5.

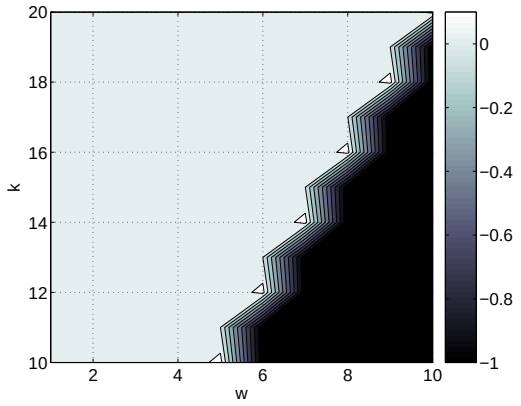


Figure 7.3: Surface plot illustrating probability bounds for a range of k and w values. For $k > 2w$, the probability is close to zero. For the very unlikely case $k \leq 2w$, the probability has been manually set to -1. Observe that the probability is much larger when $k = 2w$. This case does not satisfy the requirements of Lemma 7.6 and is rarely seen in real data.

INPUT: S, $w_{\max}$

Generate a random word β of length $w_{\max} + 1$, conditioned on $\Psi(S, \beta) > 0$.
$\alpha_1 \leftarrow \beta[w_{\max}]$
$\gamma_1 \leftarrow \beta[w_{\max} + 1]$
$p_{\mathrm{e}}(\alpha_1, \gamma_1) \leftarrow \dfrac{\Psi(S, \alpha_1)}{\Psi(S, \alpha 1.\gamma 1)}$
for $w = 2$ to $w_{\max}$ **do**
    $\alpha_w \leftarrow \beta[w_{\mathrm{wmax}} - w + 1..w_{\mathrm{wmax}}]$
    $\gamma_w \leftarrow \beta[w_{\mathrm{wmax}} - w + 2..w_{\mathrm{wmax}} + 1]$
    $p_{\mathrm{e}}(\alpha_w, \gamma_w) \leftarrow \dfrac{\Psi(S, \alpha_w \cdot \gamma_w[w])}{\Psi(S, \alpha_w)}$
    $p_{\mathrm{d}}(\alpha_w, \gamma_w) \leftarrow p_{\mathrm{e}}(\alpha_{w-1}, \gamma_{w-1})$
end for
for $w = 1$ to $w_{\max}$ **do**
    $\Delta_w \leftarrow \mathrm{L}_1(p_{\mathrm{e}}(i_w, j_w), p_{\mathrm{d}}(i_w, j_w))$
    print Δ_w
end for

Figure 7.4: Pseudocode for determining the variation of Δ_w with w.

7.4 Variation of distances between $p_{\mathrm{e}}(x, y)$ and $p_{\mathrm{d}}(x, y)$ with input sequence length

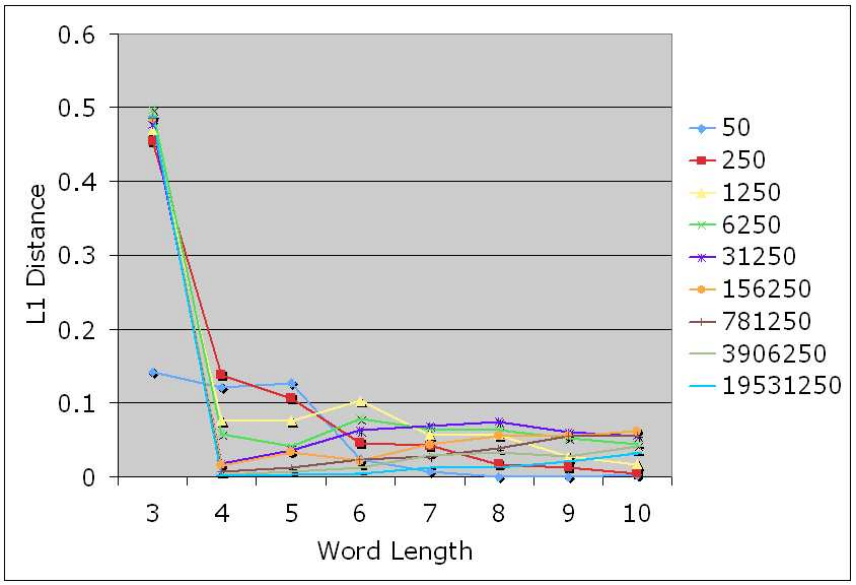
To build a Markov chain corresponding to an input sequence, enough sequence must be present for the Markov chain to sample the transitions sufficient number of times and model the input sequence correctly. This means that the input sequence should be long enough to enable sufficient sampling of each state. In the absence of a sufficiently long sequence, the Markov chain produced from the sequence is no longer a true model. Figure 7.5 examines the variation of distances between $P_{w,emp}$ and $P_{w,der}$ for binary sequences of different lengths. In Figure 7.5(a), the generating Markov chain of the sequences has order 3, while in Figure 7.5(b), the generating Markov chain of the sequences has order 5. Observe that as the sequence gets longer, the sharp transition indicating the order of the generating Markov chain becomes more and more prominent and the noise at higher word lengths decreases.

7.5 Results

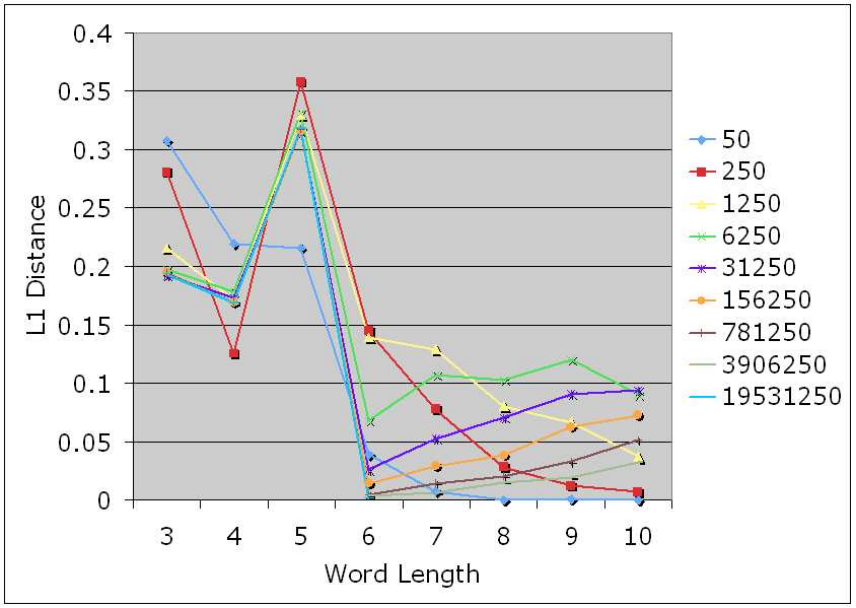
To study the ability of our algorithm to identify the order of the generating Markov chain for a given sequence, we conducted the following experiment. We randomly generated Markov chains of orders 2, 3, 4, 5, 6, and 7. Each Markov chain was used to generate sequences of length 1 Mb. Each sequence was examined at word lengths $w \in [2, 10]$.

First, we studied the behavior of Δ_w. For each sequence, β was sampled 100 times and Δ_w was computed for $w \in [2, 10]$ as described in Figure 7.4 and Section 7.3. For each w, Δ_w was plotted. Figure 7.6 illustrates the variation of Δ_w with increasing w. Observe that the change in Δ_w between consecutive values of w is negative and maximum between $w = \hat{w}$ and $w = \hat{w}+1$. Figure 7.6(a) and (b) also illustrate that $\mathbf{E}[\Delta_w] \to 0$ for $w \geq \hat{w}$. In the ideal situation, with more sequence at hand, this phenomenon is also exhibited when $\hat{w}$ is higher. Observe that as w increases, for the same sequence size, the noise increases. This can be attributed to the behavior of $\mathrm{Var}[Z]$ as described in Section 7.3. For a sequence of a given length, higher order words occur more infrequently, and their occurrence counts do not represent their real distribution. This is responsible for increased noise at higher values of w.

Figure 7.6 also suggests that it is not reliable to utilize a single instance of β to estimate $\hat{w}$. Sampling β several times gives a more reliable estimate of Δ_w values, as illustrated in Figure 7.7. For each sequence, we used 100 samples of β to compute the average value of Δ_w at each value of w. These average values are plotted in Fig 7.7. Observe that when an average over multiple samples are taken, the Δ_w curve is much



(a)



(b)

Figure 7.5: Variation of L_1 distances with input sequence length and word length variation. The x-axis indicates the word length at which the sequence is being examined. The y-axis indicates the distance between the empirical and derived distributions. Different colors indicate distance variation graphs for input sequences of different lengths. (a) An order 3 generating Markov chain was used. (b) An order 5 generating Markov chain was used.

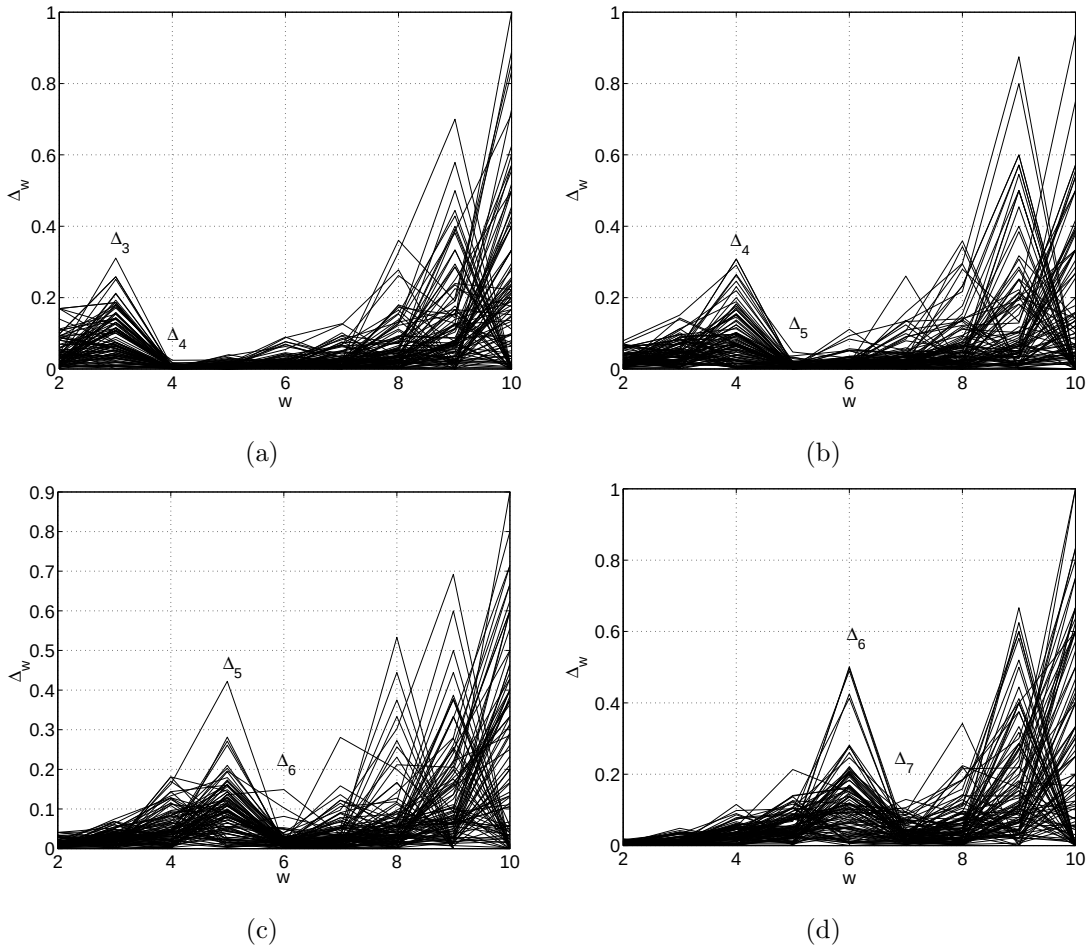


Figure 7.6: Variation of Δ_w in a sequence generated by Markov chains of different orders. (a) $\hat{w} = 3$ and (b) $\hat{w} = 4$. , (c) $\hat{w} = 5$, and (d) $\hat{w} = 6$. Observe that $\mathbf{E}[\Delta_w] \approx 0$ for $w > \hat{w}$ is demonstrated nicely by (a). This is because of the presence of ample sequence to characterize the transition probabilities at various word lengths.

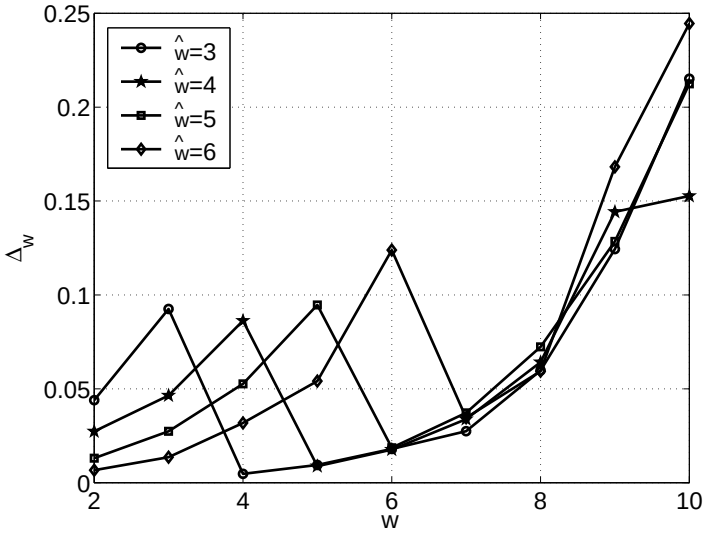


Figure 7.7: Plot of average Δ_w values over 100 samples of β. The identification of $\hat{w}$ is much more well-defined. The sharp transitions are clearer and the Δ_w curve is much smoother.

smoother and identifies $\hat{w}$ correctly in all cases.

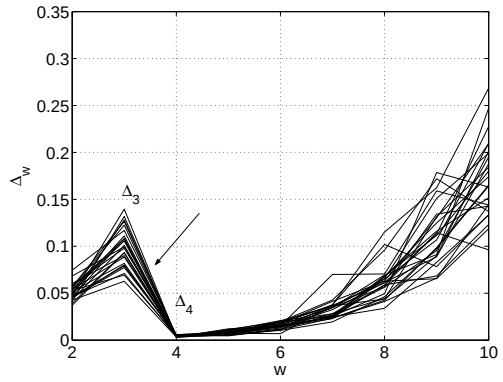
Next, we studied the effectiveness of our algorithm in identifying $\hat{w}$ for a given sequence. Using each Markov chain, we generated 25 sequences and then used average Δ_w values across multiple samples of β to estimate $\hat{w}$. Figure 7.8 illustrates the results. In 100% of the samples, our algorithm estimated $\hat{w}$ correctly. Genomic segments of *A. thaliana* including coding regions, untranslated regions, and random genomic segments of lengths 30 kilobases and 80 kilobases were studied using our estimator. Neither our estimator nor the Dalevi-Dubhashi estimator identified an order in these sequences.

7.5.1 Dependence of convergence on eigenvalues of P_w.

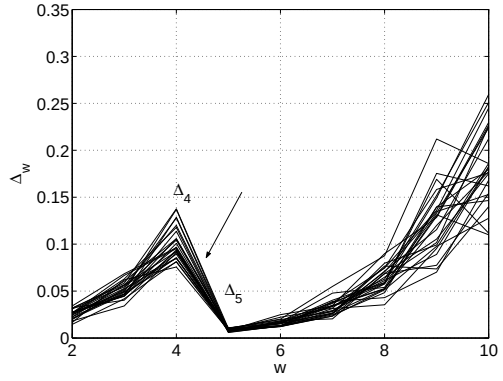
In this section, we explore the relationships between the second largest eigenvalue modulus (SLEM) of the transition matrix and the convergence of its Markov chain. We then try to relate the SLEM to the order of the generating Markov chain. A transition matrix P is said to have an eigenvalue λ if there exists a vector $v \neq 0$ such that

$$P \cdot v = \lambda \cdot v.$$

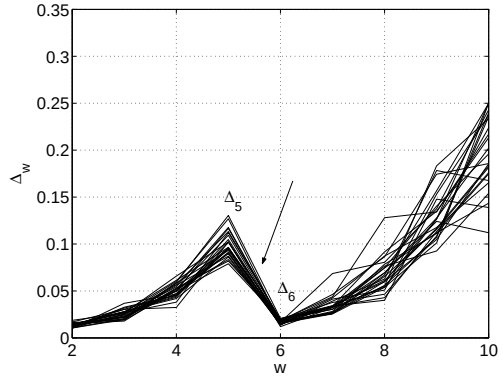
The eigenvalues of P are the roots of the characteristic polynomial $p(\lambda) = det(P - \lambda I)$, where I is the identity matrix. If $|\mathcal{S}| = n$ and P is $n \times n$, P has n eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$, and corresponding eigenvectors $v_0, v_1, \ldots, v_{n-1}$.



(a)



(b)



(c)

Figure 7.8: Effectiveness of Δ_w in identifying $\hat{w}$. 25 different sequences generated by Markov chains of order (a) 3 (b) 4 and (c) 5, respectively, have been used.

Lemma 7.7. *Any stochastic matrix P has an eigenvalue equal to 1.*

Proof. If P is a stochastic matrix, then, for the vector u such that $u(x) = 1$ for all $x \in \mathcal{S}^w$, $P \cdot u = u$. Therefore, any stochastic matrix P has an eigenvalue equal to 1. □

Let $\lambda_0 = 1$. Among $\lambda_1, \ldots, \lambda_{n-1}$, let λ_* be the largest eigenvalue. $|\lambda_*|$ is called the *Second Largest Eigenvalue Modulus (SLEM)*. Then, Lemma 7.8 holds.

Lemma 7.8. *For a stochastic matrix P, let λ_* be the SLEM of P. Then $\lambda_* \leq 1$.*

Proof. Let $Pv = \lambda v$ for some eigenvalue λ and corresponding eigenvector v. Choose $x \in \mathcal{S}^w$ such that $|v(x)| \geq |v(y)|$ for all $y \in \mathcal{S}^w$.

$$\begin{aligned} |\lambda v(x)| &= |(Pv)_x| \\ &= \left| \sum_y P(x,y)v(y) \right| \\ &\leq \sum_y |v(y)| P(x,y) \\ &\leq \sum_y |v(x)| P(x,y) \\ &\leq |v(x)|. \end{aligned}$$

So, $|\lambda| \leq 1$. Therefore, $\lambda_* \leq 1$. □

Lemma 7.9. *A finite Markov chain satisfies $\lambda_* < 1$ if and only if it is both indecomposable and aperiodic.*

A proof for Lemma 7.9 can be found in Behrends [10].

Recall that an $n \times n$ matrix P is said to be *diagonalizable* if P can be written as $P = BDB^{-1}$, where D is a diagonal $n \times n$ matrix with the eigenvalues of P as its main entries and B is an invertible (i.e., $det(B) \neq 0$) $n \times n$ matrix consisting of the eigenvectors corresponding to the eigenvalues of P.

Lemma 7.10. *If P is diagonalizable and $\lambda_* < 1$, then there is a unique stationary distribution π on $\mathcal{S}^w$. Given an initial distribution μ_0 and a point $x \in \mathcal{S}^w$,*

$$|\mu_k(x) - \pi(x)| \leq \sum_{m=1}^{n-1} |a_m v_m(x)| |\lambda_m|^k \leq \left(\sum_{m=1}^{n-1} |a_m v_m(x)| \right) (\lambda_*)^k,$$

where, k denotes the number of steps,

$$\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$$

are the eigenvalues of P,

$$v_0, v_1, \ldots, v_{n-1}$$

are a basis of the corresponding right eigenvectors, and a_m are the unique complex coefficients satisfying

$$\mu_0 = a_0 v_0 + a_1 v_1 + \ldots + a_{n-1} v_{n-1}.$$

Proof. The following proof has been taken from Behrends [10].

$$\mu_k = \mu_0 P^k. \tag{7.1}$$

$$v_m P = \lambda_m v_m. \tag{7.2}$$

$$\lambda_0 = 1. \tag{7.3}$$

Using Equations 7.1, 7.2, and 7.3, we get

$$\begin{aligned}
\mu_k &= (a_0 v_0 + a_1 v_1 + \ldots + a_{n-1} v_{n-1}) P^k && (7.4)\\
&= (a_0 v_0 + a_1 v_1 + \ldots + a_{n-1} v_{n-1})(v_m^{-1} \lambda_m v_m)^k && (7.5)\\
&= a_0 v_0 \lambda_0 + a_1 v_1 (\lambda_1)^k + \ldots + a_{n-1} v_{n-1} (\lambda_{n-1})^k. && (7.6)
\end{aligned}$$

$\lambda_* < 1$. So, $(\lambda_m)^k \to 0$ as $k \to \infty$, for $1 \leq m \leq n-1$. So, $\mu_k \to a_0 v_0$. And $a_0 = \left(\sum_y v_0(y)\right)^{-1}$. Therefore, $\pi = a_0 v_0$ is a unique probability distribution independent of μ_0 and

$$\begin{aligned}
\mu_k(x) - \pi(x) &= a_1 v_1(x)(\lambda_1)^k + \ldots + a_{n-1} v_{n-1}(x)(\lambda_{n-1})^k && (7.7)\\
\Rightarrow |\mu_k(x) - \pi(x)| &= |a_1 v_1(x)(\lambda_1)^k + \ldots + a_{n-1} v_{n-1}(x)(\lambda_{n-1})^k| && (7.8)\\
\Rightarrow |\mu_k(x) - \pi(x)| &\leq \sum_{m=1}^{n-1} |a_m v_m(x)| |\lambda_m|^k. && (7.9)
\end{aligned}$$

□

For more explanations, consult Feller [37]. The above bounds were verified by comparing to the actual number of steps needed for a Markov chain to converge. For the alphabet $\mathcal{B} = \{0, 1\}$, Markov chains of order $w = 3$ and $w = 4$ were generated. The number of steps needed for each transition matrix P to converge was computed empirically by raising P to the k^{th} power such that $norm(P^k - P^{k-1}) < \epsilon$, where $epsilon = 10^{-5}$. The theoretical bound on the number of steps was computed using Lemma 7.10 by taking the average value over all $x \in \mathcal{S}^w$. A subset of the results is summarized in Table 7.3 which demonstrates that the theoretical and empirical values of k are close.

Table 7.3: Number of steps required for matrix convergence. Comparison of number of steps required for transition matrix convergence, when computed empirically and theoretically.

w	**k** (empirical)	**k** (theoretical)
3	18	18.375
3	44	47.750
3	53	56.625
3	32	34.000
4	74	73.500
4	84	92.937
4	44	43.000
4	33	34.313

The following experiment is done to study the SLEMs and steps needed for convergence of Markov chains of various orders computed from a sequence generated by a Markov chain of fixed order. Markov chains of orders 3 until 10 are generated by randomly assigning probabilities to their transition matrices, while maintaining their stochastic nature. These Markov chains are used to generate sequences.

Proposition 1. *If a sequence S has a generating Markov chain of order $\hat{w}$, then, the SLEMs of the transition matrices of all Markov chains of order $\hat{w}+1$ and greater generated from S will differ by ϵ, where $\epsilon \to 0$, as $S \to \infty$.*

While we do not have a theoretical proof for the above proposition yet, Figure 7.9 illustrates values that agree with Proposition 1. Figure 7.10 illustrates the number of steps needed for convergence in generated Markov chains. We observe that while the SLEM varies very little for $w > \hat{w}+1$, the number of steps required for convergence does not stabilize for generated Markov chains of orders w and greater.

Further exploration in this direction is one of the future directions of research we are exploring.

7.6 Conclusions and Future work

In this work, we have built a formal framework for the analysis of sequences using DNA words of different lengths and illustrated the performance of the algorithm we use to estimate Markov chain order. With suitable sampling, it is possible to predict the order of generating Markov chains using much shorter sequences. The exact method and the corresponding mathematical framework is one of the directions we are pursuing.

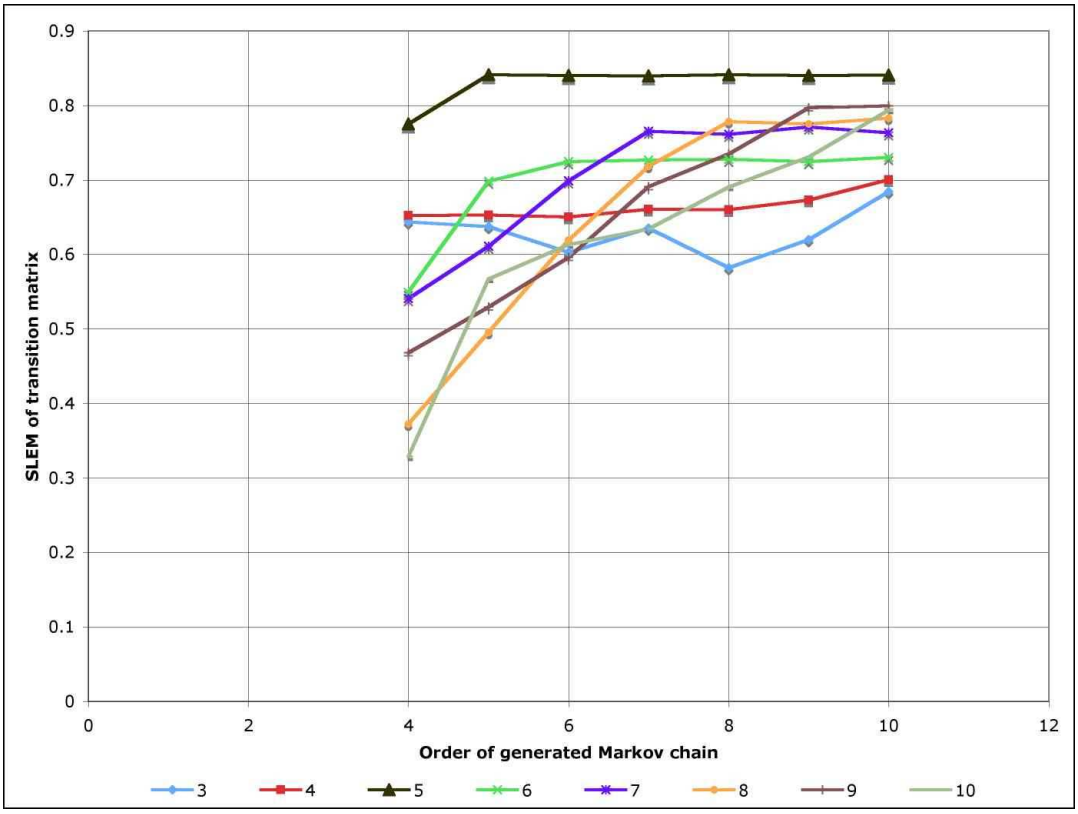


Figure 7.9: Variation of SLEMs in *generated* Markov chains of different orders. Different colors indicate *generated* transition matrix SLEM trends for *generating* Markov chains of different orders. The color legend gives the *generating* Markov chain orders.

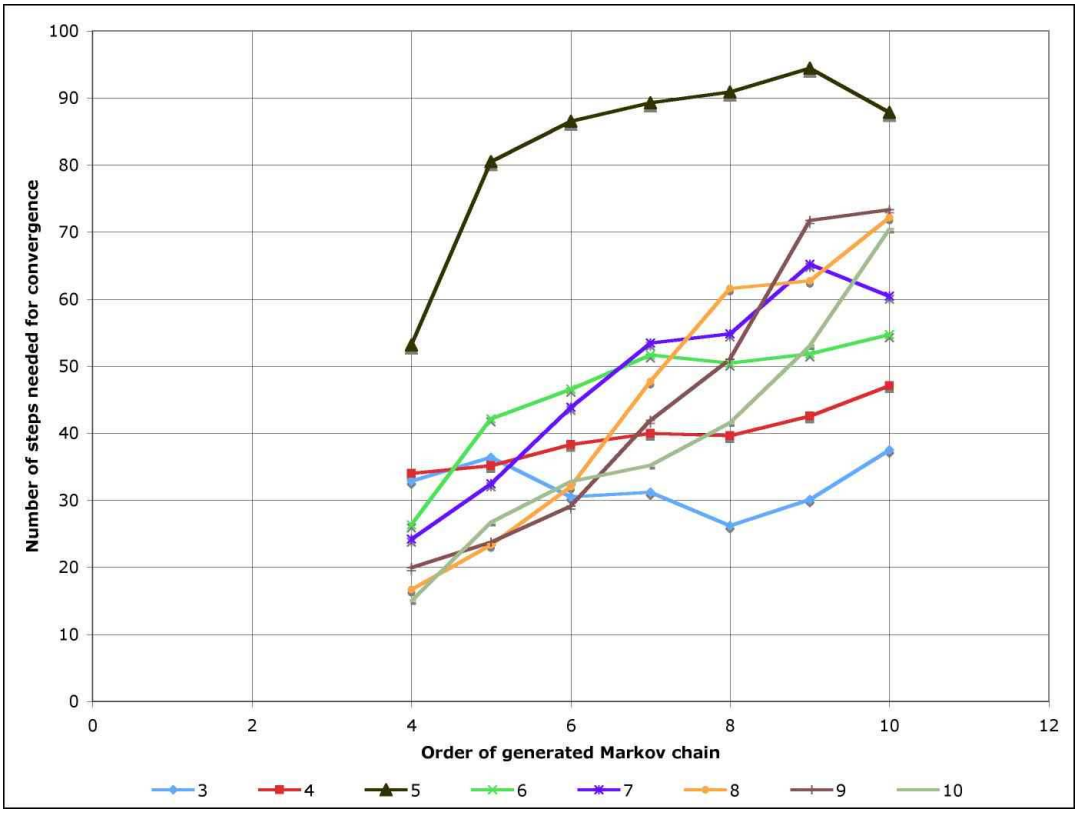


Figure 7.10: Variation of steps needed for convergence in *generated* Markov chains of different orders. Different colors indicate steps required for *generated* transition matrix convergence for *generating* Markov chains of different orders. The color legend gives the *generating* Markov chain orders.

Comparison of the performance of our algorithm to that of the Peres-Shields and Dalevi-Dubhashi estimators is also one of the future directions. Having sufficient sequence to summarize the transition probabilities accurately at all word lengths is also important. Ultimately, the amount of available sequence, the range of word lengths, the behavior of the variance, and the transition probabilities, can all be integrated to compute an efficient Markov order estimator from sequence.

Chapter 8

Conclusions

Genomic signatures computed from sequences that use oligonucleotide frequencies have been studied extensively in scientific literature. In computing the de Bruijn chain signature, we have integrated aspects of graph structure and Markov chain stationary distributions to extract unique aspects of genomic sequences. Both the stationary distribution and graph-based signatures are novel approaches that have not been explored previously in the scientific literature. The graph-based component of the DBC signature is a measure of the strength of connectivity of each vertex to the largest connected component of the graph of which it is part, while the stationary distribution is a relative of word frequency based signatures that characterizes the underlying Markov chain more closely. Together, these two features result in a powerful signature that has a better accuracy of origin prediction for short DNA sequences than existing word frequency based signatures.

Using a collection of species sampled uniformly from all parts of the taxonomic tree, we have demonstrated that the θ_2^{dbc} signature is very well-conserved in all species except the set of tetrapod vertebrates in our collection. We demonstrate that the θ_2^{dbc} signature is able to accurately distinguish between diverse species as well as closely-related species.

In this work, besides exploring the properties of the θ_2^{dbc} signature using empirical results, we build a theoretical framework within which we characterize the separation between θ_2^{dbc} signatures of DNA fragments hypothesized to be generated by either the same or different de Bruijn chains. We obtain probabilistic bounds on separation using parameters of the hypothetical generating de Bruijn chain(s). Additionally, we also establish a mathematical framework for the word count vector, which is novel.

Several interesting computational problems arise from the study of genomic signatures. Distances between θ_2^{dbc} signatures can serve as a basis for phylogenetic reconstruction. This would eliminate the need for

computation-intensive alignments. Another possible direction is to study the conservation of each genomic signature under every level of organization in the taxonomic tree. This will help identify the extent to which the organisms under every level of organization are related to each other. Tree-wide analysis will also help to identify horizontal transfers between species. Moreover, alternate positions for organisms in the taxonomic tree may be identified. A large-scale software system that stores the genomic signatures of all sequenced genomes and is meant to identify the origin and close relatives of short segments of DNA is under construction.

Genomic signatures are the central topic of this dissertation. Two other computational problems have also been addressed in this work. The first problem deals with the estimation of Markov chain order. Given a sequence hypothesized to be generated by a Markov chain, we propose a mathematical framework and an algorithm to estimate the order of that Markov chain using properties of oligonucleotides in the sequence. The second topic is a part of the Computational Models for Gene Silencing project. It consists of a centralized database for all types of biological data for *C. elegans*. Associated computational tools perform data-mining operations on this data and enrich the database with the computed results. Raw data as well as hypotheses generated by the data-mining methods are served using an associated website.

REFERENCES

[1] K. Afshar, F. S. Willard, K. Colombo, D. P. Siderovski, and P. Gonczy. Cortical localization of the G-alpha protein GPA-16 requires RIC-8 function during *C. elegans* asymmetric cell division. *Development*, 132:4449–4459, 2005.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 11:5–33, 2005.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, 19:716–723, 1974.

[4] David Aldous, Geoffrey R. Grimmett, C. Douglas Howard, Harry Kesten, Fabio Martinelli, Laurent Saloff-Coste, and J. Michael Steele. *Probability on Discrete Structures*, volume 110. Springer-Verlag, 2000.

[5] Arnold O. Allen. *Probability, Statistics, and Queueing Theory With Computer Science Applications.* Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, second edition, 1990.

[6] C. C. Mello amd B. W. Draper and J. R. Priess. The maternal genes apx-1 and glp-1 and establishment of dorsal-ventral polarity in the early C. elegans embryo. *Cell*, 77:95–106, 1994.

[7] Ariel Anguiano and Anil Potti. Genomic signatures individualize therapeutic decisions in non-small-cell lung cancer. *Future Drugs*, 7(6):837–844, 2007.

[8] C. A. Ball, I. A. Awad, J. Demeter, J. Gollub, J. M. Hebert, T. Hernandez-Boussard, H. Jin, J. C. Matese, M. Nitzberg, and F. Wymore et al. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research*, 33:D580–582, 2004.

[9] S. E. Basham and L. S. Rose. Mutations in ooc-5 and ooc-3 disrupt oocyte formation and the reestablishment of asymmetric PAR protein localization in two-cell *Caenorhabditis elegans* embryos. *Dev Biol*, 215:253–263, 1999.

[10] Ehrhard Behrends. *Introduction to Markov Chains with Special Emphasis on Rapid Mixing.* Advanced Lectures in Mathematics. Friedr. Vieweg & Sohn, Braunschweig, 2000.

[11] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67, 2002.

[12] L. A. Berkowitz and S. Strome. MES-1, a protein required for unequal divisions of the germline in early C. elegans embryos, resembles receptor tyrosine kinases and is localized to the boundary between the germline and gut cells. *Development*, 127:4419–4431, 2000.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[14] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing Markov chain on a graph. *SIAM Review*, 46(4):667–689 (electronic), 2004.

[15] Andrew V. Z. Brower. Problems with DNA barcodes for species delimitation: Ten species of *Astraptes fulgerator* reassessed (Lepidoptera: Hesperiidae). *Systematics and Biodiversity*, 4(2):127–132, 2006.

[16] A. M. Campbell, J. Mrazek, and S. Karlin. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *PNAS*, 96:9184–9189, 1999.

[17] C. H. Cannon, C. S. Kua, E. K. Lobenhofer, and P. Hurban. Capturing genomic signatures of dna sequence variation using a standard anonymous microarray platform. *Nucleic Acids Research*, 34(18), 2006.

[18] A. Carbone, F. Kepes, and A. Zinovyev. Codon bias signatures, organization of micro-organisms in codon space, and lifestyle. *Molecular Biology and Evolution*, 22(3):547–561, 2005.

[19] Jeffrey T. Chang and Joseph R. Nevins. GATHER: A systems approach to interpreting genomic signatures. *Bioinformatics*, 22(23):2926–2933, 2006.

[20] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[21] S. Cho, K. W. Rogers, and D. S. Fay. The C. elegans glycopeptide hormone receptor ortholog, FSHR-1, regulates germline differentiation and survival. *Current Biology*, 17:203–212, 2001.

[22] T. Coenye and P. Vandamme. Use of the genomic signature in bacterial classification and identification. *Systematic and Applied Microbiology*, 27(2):175–185, 2004.

[23] E. J. Cram, H. Shang, and J. E. Schwarzbauer. A systematic RNA interference screen reveals a cell migration gene network in C. elegans. *J Cell Sci*, 119:4811–4818, 2006.

[24] S. L. Crittenden, D. Rudel, J. Binder, T. C. Evans, and J. Kimble. Genes required for GLP-1 asymmetry in the early *Caenorhabditis elegans* embryo. *Dev Biol*, 181:36–46, 1997.

[25] Imre Csiszár and Paal C. Shields. The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28(6):1601–1619, 2000.

[26] S. P. Curran and G. Ruvkun. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet*, 3:e56, 2007.

[27] D. Dalevi, D. Dubhashi, and M. Hermansson. Bayesian classifiers for detecting HGT using fixed and variable order Markov models of genomic signatures. *Bioinformatics*, 22(5):517–522, 2006.

[28] Daniel Dalevi and Devdatt Dubhashi. The Peres-Shields order estimator for fixed and variable length markov chains with applications to DNA sequence similarity. In *Algorithms in Bioinformatics*, Lecture Notes in Computer Science/Lecture Notes in Bioinformatics, Mallorca, Spain, October 2005. Springer-Verlag.

[29] P. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, and B. Fertil. GENSTYLE. Website.

[30] P. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, and B. Fertil. Genomic signature is preserved in short DNA fragments. In *Proceedings of the 1st IEEE International Symposium on Bioinformatics and Biomedical Engineering*, 2000.

[31] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: Characterization and classification of species assessed by Chaos Game Representation of sequences. *Molecular Biology and Evolution*, 16(10):1391–1399, 1999.

[32] Holly K. Dressman, Andrea Bild, Jennifer Garst, David Harpole Jr, and Anil Potti. Genomic signatures in non-small-cell lung cancer: Targeting the targeted therapies. *Current Oncology Reports*, 8(4):252–257, 2006.

[33] C. Dufraigne, B. Fertil, S. Lespinats, A. Giron, and P. Deschavanne. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research*, 33(1):12 pages, 2005.

[34] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 2002.

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

[36] T-H Fan and C Tsai. A bayesian method in determining the order of a finite state Markov chain. *Statitical Theory and Methods*, 28(7):1711–1730, 1999.

[37] William Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons Inc., New York, third edition, 1968.

[38] Bernard Fertil, Matthieu Massin, Sylvain Lespinats, Caroline Devic, Philippe Dumee, and Alain Giron. GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Research*, 33 (Web Server issue):W512–W515, 2005.

[39] J. W. Fickett, D. C. Torney, and D. R. Wolf. Base compositional structure of genomes. *Genomics*, 13(4):1056–1064, 1992.

[40] A Fire, S Xu, MK Montgomery, SA Kostas, SE Driver, and CC Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391:806–811, 1998.

[41] A. R. Frand, S. Russel, and G. Ruvkun. Functional genomic analysis of C. elegans molting. *PLoS Biol*, 3:e312, 2005.

[42] Andrew J. Gentles and Samuel Karlin. Genome-scale compositional comparisons in eukaryotes. *Genome Research*, 11:540–546, 2001.

[43] L. R. Girard, T. J. Fiedler, T. W. Harris, F. Carvalho, I. Antoshechkin, M. Han, P. W. Sternberg, L. D. Stein, and M. Chalfie. WormBook: The online review of *Caenorhabditis elegans* biology. *Nucleic Acids Res*, 35:D472–475, 2007.

[44] Nick Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Research*, 21(10):2487–2491, 1993.

[45] J. A. Govindan, H. Cheng, J. E. Harris, and D. Greenstein. G-alpha o/i and G-alpha s signaling function in parallel with the MSP/Eph receptor to control meiotic diapause in C. elegans. *Current Biology*, 16:1257–1268, 2006.

[46] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, third edition, 2001.

[47] P. Groth, N. Pavlova, I. Kalev, S. Tonov, G. Georgiev, H. D. Pohlenz, and B. Weiss. PhenomicDB: A new cross-species genotype/phenotype resource. *Nucleic Acids Research*, 35:D696–D699, 2005.

[48] G. Grothaus, Adeel Mufti, and T.M. Murali. Automatic layout and visualization of biclusters. *Algorithms for Molecular Biology*, 2006.

[49] B. Hamilton, Y. Dong, M. Shindo, W. Liu, I. Odell, G. Ruvkun, and S. S. Lee. A systematic RNAi screen for longevity genes in C. elegans. *Genes Dev*, 19:1544–1555, 2005.

[50] M. Prakash Hande, Tamara V. Azizova, Charles R. Geard, Ludmilla E. Burak, Catherine R. Mitchell, Valentin F. Khokhryakov, Evgeny K. Vasilenko, and David J. Brenner. Past exposure to densely ionizing radiation leaves a unique permanent signature in the genome. *The American Society of Human Genetics*, 72:1162–1170, 2003.

[51] M. Hansen, S. Taubert, D. Crawford, N. Libina, S. J. Lee, and C. Kenyon. Lifespan extension by conditions that inhibit translation in *Caenorhabditis elegans*. *Aging Cell*, 6:95–110, 2007.

[52] Lenwood S. Heath and Amrita Pati. Genomic signatures from DNA word graphs. In *Lecture Notes in Bioinformatics*, volume 4463, pages 317–328. Springer-Verlag, 2007.

[53] Lenwood S. Heath and Amrita Pati. Genomic signatures in de Bruijn chains. In *Lecture Notes in Bioinformatics: Algorithms in Bioinformatics (WABI 2007)*, volume 4645, pages 216–227. Springer-Verlag, 2007.

[54] Lenwood S. Heath and Amrita Pati. Predicting Markov chain order in genomic sequences. In *BIBM 2007*, pages 159–164. IEEE Computer Society, 2007.

[55] Paul D. N. Hebert, Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. Biological identifications through DNA barcodes. *Proceedings of the Royak Society, Biological Science*, 270(1512):313–321, 2003.

[56] Paul D. N. Hebert, Erin H. Penton, John M. Burns, Daniel H. Janzen, and Winnie Hallwachs. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *PNAS*, 101(41):14812–14817, 2004.

[57] Paul D. N. Hebert, Mark Y. Stoeckle, Tyler S. Zemlak, and Charles M. Francis. Identification of birds through DNA barcodes. *PLoS Biology*, 2(10):e312, 2004.

[58] Kathleen A. Hill, Nicholas J. Schisler, and Shiva M. Singh. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *Journal of Molecular Evolution*, 35(3):261–269, 1992.

[59] Paul G. Hoel, Sidney C. Port, and Charles J. Stone. *Introduction to Stochastic Processes.* Houghton Mifflin Company, 1972.

[60] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Publishing Co., Reading, Mass., 1979. Addison-Wesley Series in Computer Science.

[61] John M. Howie. *Fundamentals of Semigroup Theory*. Clarendon Press, Oxford, 1995.

[62] C. J. Hsiao and M. J. Zaki. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, 17:462–478, 2005.

[63] H. Joel Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.

[64] R. W. Jernigan and R. H. Baran. Pervasive properties of the genomic signature. *BMC Genomics*, 3:9 pages, 2002.

[65] Mao Jiang-Hua, Jiangzhen Li, Tao Jiang, Qian Li, Di Wu, Jesus Perez-Losada, Reyno DelRosario, Leif Peterson, Wei-Wen Cai, and Allan Balmain. Genomic signatures for radiation-induced mouse lymphoma. Available on: `http://www.lowdose.energy.gov`, 2006.

[66] Y. Jin, T. M. Murali, and N. Ramakrishnan. Compositional mining of multi-relational biological datasets. *ACM Transactions on Knowledge Discovery from Data*, 2008.

[67] Pal Kaposi-Novak, Ju-Seog Lee, Luis Gmez-Quiroz, Cédric Coulouarn, Valentina M. Factor, and Snorri S. Thorgeirsson. Met-regulated expression signature defines a subset of human hepatocellular carcinomas with poor prognosis and aggressive phenotype. *Journal of Clinical Investigation*, 116:1582–1595, 2006.

[68] S. Karlin and C. Burge. Dinucleotide relative abundance extremes — A genomic signature. *Trends in Genetics*, 11(7):283–290, 1995.

[69] S. Karlin, I. Landunga, and B.E. Blaisdell. Heterogeneity of genomes: Measures and values. *PNAS*, 91:12837–12841, 1994.

[70] S. Karlin, J. Mrazek, and A. M. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, 179(12):3899–3913, 1997.

[71] W. John Kress, Kenneth J. Wurdack, Elizabeth A. Zimmer, Lee A. Weigt, and Daniel H. Janzen. Use of DNA barcodes to identify flowering plants. *PNAS*, 102(23):8369–8374, 2005.

[72] Renaud Lahaye, Michelle van der Bank, Diego Bogarin, Jorge Warner, Franco Pupulin, Guillaume Gigot, Olivier Maurin, Sylvie Duthoit, Timothy G. Barraclough, and Vincent Savolainen. DNA barcoding the floras of biodiversity hotspots. *PNAS*, 2008.

[73] D. M. Lambert, A. Baker, L. Huynen, O. Haddrath, P. D. N. Hebert, and C. D. Millar. Is a large-scale DNA-based inventory of ancient life possible? *Journal of Heredity*, 96(3):279–284, 2005.

[74] T. Lamitina, C. G. Huang, and K. Strange. Genome-wide RNAi screening identifies protein damage as a regulator of osmoprotective gene expression. *PNAS*, 103:12173–12178, 2006.

[75] J. Y. Lee, D. J. Marston, T. Walston, J. Hardin, A. Halberstadt, and B. Goldstein. Wnt/Frizzled signaling controls C. elegans gastrulation by activating actomyosin contractility. *Curr Biol*, 16:1986–1997, 2005.

[76] G. Lettre, E. A. Kritikou, M. Jaeggi, A. Calixto, A. G. Fraser, R. S. Kamath, J. Ahringer, and M. O. Hengartner. Genome-wide RNAi identifies p53-dependent and -independent regulators of germ cell apoptosis in C. elegans. *Cell Death Differ*, 11:1198–1203, 2004.

[77] Harry R. Lewis and Christos H. Papadimitriou. *Element of the Theory of Computation*. Prentice-Hall, Upper Saddle River, New Jersey, second edition, 1998.

[78] R. Lin, S. Thompson, and J. R. Priess. pop-1 encodes an HMG box protein required for the specification of a mesoderm precursor in early C. elegans embryos. *Cell*, 83:599–609, 1995.

[79] K Liolios, K Mavrommatis, N Tavernarakis, and NC Kyrpides. The Genomes On Line Database (GOLD) in 2007: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 36:D475–D479, 2008.

[80] M. C. Lo, F. Gay, R. Odo, Y. Shi, and R. Lin. Phosphorylation by the beta-catenin/MAPK complex promotes 14-3-3-mediated nuclear export of TCF/POP-1 in signal-responsive cells in C. elegans. *Cell*, 117:443–453, 2004.

[81] D.R. Maddison and K.-S. Schulz. The Tree of Life web project. Available on: `http://tolweb.org`, 2007.

[82] S. C. Madeira and A. L. Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:1311–1316, 2004.

[83] I. Maeda, Y. Kohara, M. Yamamoto, and A. Sugimoto. Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Nucleic Acids Research*, 11:171–176, 2001.

[84] Susanna Mandruzzato, Andrea Callegaro, Gianluca Turcatel, Samuela Francescato, Maria C Montesco, Vanna Chiarion-Sileni, Simone Mocellin, Carlo R Rossi, Silvio Bicciato, Ena Wang, Francesco M Marincola, and Paola Zanovello. A gene expression signature associated with survival in metastatic melanoma. *Journal of Translational Medicine*, 4(50), 2006.

[85] P. Mendiratta and P.G. Febbo. Genomic signatures associated with the development, progression, and outcome of prostate cancer. *Molecular Diagnosis and Therapy*, 11(6):345–354, 2007.

[86] Christopher P. Meyer and Gustav Paulay. DNA Barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, 3(12):e422, 2005.

[87] Michael Mitzenmacher and Eli Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[88] D. G. Morton, J. M. Roos, and K. J. Kemphues. par-4, a gene required for cytoplasmic localization and determination of specific cell types in *Caenorhabditis elegans* embryogenesis. *Genetics*, 130:771–790, 1992.

[89] Hiroshi Nakashima, Ken Nishikawa, and Tatsuo Ooi. Differences in dinucleotide frequencies of human, yeast, and *Escherichia coli* Genes. *DNA Research*, 4:185–192, 1997.

[90] Hiroshi Nakashima, Motonori Ota, Ken Nishikawa, and Tatsuo Ooi. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Research*, 5:251–259, 1998.

[91] A. Nanopoulos and Y. Manolopoulos. Efficient similarity search for market basket data. *VLDB Journal*, 11:138–152, 2002.

[92] E. A. Nollen, S. M. Garcia, G. van Haaften, S. Kim, A. Chavez, R. I. Morimoto, and R. H. Plasterk. Genome-wide RNA interference screen identifies previously undescribed regulators of polyglutamine aggregation. *PNAS*, 101:6403–6408, 2004.

[93] Benjamin B. Normark, Olivia P. Judson, and Nancy A. Moran. Genomic signatures of ancient asexual lineages. *Biological Journal of the Linnean Society*, 79:69–84, 2003.

[94] J.L. Oliver, P. Bernaola-Galván, J. Guerrero-García, and R. Román-Roldán. Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretical Biology*, 160(4):457–470, 1993.

[95] Yuval Peres and Paul Shields. Two new Markov order estimators. *ArXiv Mathematics e-prints*, June 2005.

[96] P. A. Pevzner. DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica*, 13(1-2):77–105, 1995.

[97] P. A. Pevzner, H. X. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *PNAS*, 98(17):9748–9753, 2001.

[98] J. Pothof, G. van Haaften, K. Thijssen, R. S. Kamath, A. G. Fraser, J. Ahringer, R. H. Plasterk, and M. Tijsterman. Identification of genes that protect the C. elegans genome against mutations by genome-wide RNAi. *Genes Dev*, 17:443–448, 2003.

[99] Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster, and Joseph R Nevins. Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, 12:1294–1300, 2006.

[100] G. Poulin, Y. Dong, A. G. Fraser, N. A. Hopper, and J. Ahringer. Chromatin regulation and sumoylation in the inhibition of Ras-induced vulval development in *Caenorhabditis elegans*. *Embo J*, 24:2613–2623, 2005.

[101] B. Raphael, D. G. Zhi, H. X. Tang, and P. Pevzner. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Research*, 14(11):2336–2346, 2004.

[102] Sujeevan Ratnasingham and Paul D. N. Hebert. BOLD: The barcode of life data system. *Molecular Ecology Notes*, 2007.

[103] Arnold L. Rosenberg and Lenwood S. Heath. *Graph Separators, With Applications*. Frontiers of Computer Science. Kluwer Academic/Plenum Publishers, 2000.

[104] Jeffrey S. Rosenthal. Convergence rates for Markov chains. *SIAM Review*, 37(3):387–405, September 1995.

[105] M. C. Saleh, R. P. van Rij, A. Hekele, A. Gillis, E. Foley, P. H. O'Farrell, and R. Andino. The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. *Nat Cell Biol*, 8:793–802, 2006.

[106] R. Sandberg, C. I. Branden, I. Ernberg, and J. Coster. Quantifying the species-specificity in genomics signatures, synonymous codon choice, amino acid usage, and G+C content. *Gene*, 311:35–42, 2003.

[107] S. Sarawagi and A. Kirpal. Efficient set joins on similarity predicates. *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'04)*, pages 743–754, 2002.

[108] Christoffer Schander and Endre Willassen. What can biological barcoding do for marine biology? *Marine Biology Research*, 1(1):79–83, 2005.

[109] C. Schmitz, P. Kinge, and H. Hutter. Axon guidance genes identified in a large-scale RNAi screen using the RNAi-hypersensitive *Caenorhabditis elegans* strain nre-1(hd20) lin-15b(hd126). *PNAS*, 104:834–839, 2007.

[110] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[111] D. Sieburth, Q. Chong, M. Dybbs, M. Tavazoie, S. Kennedy, D. Wang, D. Dupuy, J. F. Rual, D. E. Hill, and M. Vidal et al. Systematic analysis of genes required for synapse structure and function. *Nature*, 436:510–517, 2005.

[112] K. R. Siegfried and J. Kimble. POP-1 controls axis formation during early gonadogenesis in C. elegans. Development. *Cell*, 129:443–453, 2002.

[113] M. Alex Smith, D. Monty Wood, Daniel H. Janzen, Winnie Hallwachs, and Paul D. N. Hebert. DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *PNAS*, 104(12):4967–4972, 2007.

[114] M. Alex Smith, Norman E. Woodley, Daniel H. Janzen, Winnie Hallwachs, and Paul D. N. Hebert. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *PNAS*, 103(10):3657–3662, 2006.

[115] G. W. Snedecor and W. G. Cochran. *The Sample Correlation Coefficient r and Properties of r*. Iowa State Press, 7^{th} edition, 1980.

[116] Eileen Solan and Nicolas Vieille. Perturbed Markov chains. *Journal of Applied Probability*, 40:107–122, 2003.

[117] B. Sonnichsen, L. B. Koski, A. Walsh, P. Marschall, B. Neumann, M. Brehm, A. M. Alleaume, J. Artelt, P. Bettencourt, and E. Cassin et al. Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature*, 434:462–469, 2005.

[118] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34:D535, 2004.

[119] K. K. Stein, E. S. Davis, T. Hays, and A. Golden. Components of the spindle assembly checkpoint regulate the anaphase-promoting complex during meiosis in *Caenorhabditis elegans*. *Genetics*, 175:107–123, 2005.

[120] Y. Suzuki and M. Han. Genetic redundancy masks diverse functions of the tumor suppressor gene PTEN during C. elegans development. *Genes Dev*, 20:423–428, 2006.

[121] J. Symersky, Y. Zhang, N. Schormann, S. Li, R. Bunzel, P. Pruett, C. H. Luan, and M. Luo. Structural genomics of *Caenorhabditis elegans*: Structure of the BAG domain. *Acta Crystallogr D Biol Crystallogr*, 60:1606–1610, 2004.

[122] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:S136–S144, 2002.

[123] H. Teeling, A. Meyerdierks, M. Buaer, R. Amann, and F. O. Glockner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6:938–947, 2004.

[124] V. Urquidia and S. Goodison. Genomic signatures of breast cancer metastasis. *Cytogenetic and Genome Research*, 118:116–129, 2007.

[125] G. van Haaften, N. L. Vastenhouw, E. A. Nollen, R. H. Plasterk, and M. Tijsterman. Gene interactions in the DNA damage-response pathway identified by genome-wide RNA-interference analysis of synthetic lethality. *PNAS*, 101:12992–12996, 2004.

[126] M. W. J. van Passel, A. Bart, H. H. Thygesen, A. C. M. Luyf, A. H. C. van Kampen, and A. van der Ende. An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics*, 6:10 pages, 2005.

[127] Mark W. J. van Passel, Eiko E Kuramae, Angela C. M. Luyf, Aldert Bart, and Teun Boekhout. The reach of the genome signature in prokaryotes. *BMC Evolutionary Biology*, 6(84):8 pages, 2006.

[128] N. L. Vastenhouw, S. E. Fischer, V. J. Robert, K. L. Thijssen, A. G. Fraser, R. S. Kamath, J. Ahringer, and R. H. Plasterk. A genome-wide screen identifies 27 genes involved in transposon silencing in C. elegans. *PNAS*, 13:1311–1316, 2003.

[129] Y. W. Wang, K. Hill, S. Singh, and L. Kari. The spectrum of genomic signatures: From dinucleotides to chaos game representation. *Gene*, 346:173–185, 2005.

[130] Michael Waterman. *Introduction to Computational Biology*. Academic Press Inc., Boston, MA, first edition, 1995.

[131] J. L. Watts, B. Etemad-Moghadam, S. Guo, L. Boyd, B. W. Draper, C. C. Mello, J. R. Priess, and K. J. Kemphues. par-6, a gene involved in the establishment of asymmetry in early C. elegans embryos, mediates the asymmetric localization of PAR-3. *Development*, 122:3133–3140, 1996.

[132] Martin Wiemers and Konrad Fiedler. Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, 4(8), 2007.

[133] M. J. Zaki and C. J. Hsiao. CHARM: An efficient algorithm for closed itemset mining. *SIAM International Conference on Data Mining*, pages 457–473, 2002.

[134] Y. Zhang and M. S. Waterman. An Eulerian path approach to global multiple alignment for DNA sequences. *Journal of Computational Biology*, 10(6):803–819, 2003.

[135] Y. Zhang and M. S. Waterman. An Eulerian path approach to local multiple alignment for DNA sequences. *PNAS*, 102(5):1285–1290, 2005.

[136] W. Zhong and P. W. Sternberg. Genome-wide prediction of C. elegans genetic interactions. *Science*, 311:1481–1484, 2006.

Appendix A

CMGSDB: Integrating heterogeneous *C. elegans* data sources using compositional data mining

Other contributors to this work are Ying Jin, Karsten Klage, Lenwood S. Heath, Richard Helm, and Naren Ramakrishnan.

Introduction

The availability of high-throughput screens has opened up awareness of the importance of data integration to reveal useful biological insight. For instance, the study of even a focused aspect of cellular activity, such as gene action, now benefits from multiple high-throughput data acquisition technologies, such as microarrays, genome-wide deletion screens, and RNAi assays. While enormous quantities of data are available, it remains a major challenge to construe meaningful biological evidence from this data that explains, for example, the role of a biological pathway, the effects of a SNP on disease phenotypes, or the regulatory networks or metabolic pathways underlying a cellular state. Two major factors make this process harder. First, high-throughput experiments for a given genome are performed by independent groups of researchers that develop their own naming conventions and schemes for information storage and retrieval. This makes it difficult for scientists to utilize all available data for a genome to draw inferences. Second, even if such integration is accomplished, the possibility of linking data across sources is often restricted to individual entities, such as

genes or proteins; it is difficult to track sets of entities, which is the more natural way to interact with such databases.

As a case in point, consider the possibilities of integration opened up by the availability of RNAi screens. Post-transcriptional gene silencing via RNAi was first described in the nematode Caenorhabditis elegans (*C. elegans*, CE) [40], and is presently utilized for a variety of functional genomics experiments using RNAi assays. Although Wormbase serves as a centralized repository for *C. elegans* data, the sources of RNAi experiments in *C. elegans* are many, their data representation formats are varied, and some information is lost while integrating them into the Wormbase [43] schema.

Here, we present CMGSDB, a database for computational models in gene silencing, where the following goals have been achieved. We have integrated genome annotation data, gene expression data, protein interaction data, gene regulation data, GO (Gene Ontology) annotation data, and RNAi data for *C. elegans* into a centralized schema. RNAi experiments and phenotypes have been integrated from independent research groups into a single schema. A common hierarchical structure has been designed to organize the phenotypes from different sources. The hierarchy is accessible via a web browser. Compositional data mining [66] is used to identify relationships among sets of entities across the database schema, where these sets are mined automatically and not defined *a priori*. A detailed web interface that reports all the data and the patterns computed is available at `https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/`.

Compositional Data Mining

The basic idea in compositional data mining is to mirror the shift-of-vocabulary as we traverse a database schema in a composition of data mining algorithms that mine the respective entities and relationships. For instance, consider a multiple stress environment where numerous physiological responses are occurring simultaneously. Efforts to identify a set of *C. elegans* genes (perhaps encoding transcription factors (TFs)) to knock down (via RNAi) in order to ascertain key mechanisms of response might begin by identifying those genes whose knock down produces phenotypes that modulate survival, and then find one or more transcription factors that combinatorially control the expression of these genes. This analysis can be modeled as a chain: transcription factors $\rightarrow$ genes $\rightarrow$ phenotypes. Each step in this chain is computed using a data mining algorithm, so that we first mine the relationship between transcription factors and genes for concerted (TF, gene) sets called *biclusters*, then mine the relationship between genes and phenotypes to find concerted biclusters of (gene, phenotype) pairs. The biclusters share the gene boundary leading us to investigate if these biclusters approximately match at the gene interface. The projection of the biclusters with an approximate match at one interface is called a *redescription*. Thus, compositional data mining is a way of

problem decomposition (see [66] for more details) where biclustering and redescription mining algorithms are chained in a way that mirrors the underlying "join-order" path in the database schema.

As illustrated in Figure A.1, we mine biclusters between genes and the transcription factors that regulate them, mine biclusters between genes and the phenotypes that result when they are knocked down, and relate one side of the first bicluster with one side of the second bicluster. Hence the task of integrating diverse data sources is reduced to composing data mining patterns computed over each of the sources separately. The advantage of this formulation is that each data source can be mined individually using a biclustering algorithm that is suited for that purpose. For instance, the xMotif [48], SAMBA [122], and ISA [11] algorithms are suited for mining numeric data (e.g., such as gene expression relationships), while Apriori [2] and CHARM [62] algorithms are suited for mining boolean data (e.g., graph adjacencies).

The approximate matching of biclusters is ensured using a similarity search algorithm or redescription mining approach. This problem, in various guises, has been studied by the database community; see [91] and [107] for examples. In this work, we utilize a cover-tree approach for fast computation of similar biclusters. The overlap between the sides of biclusters is qualified using the Jaccard's coefficient: the Jaccard's coefficient between two sets X and Y is the ratio

$$|X \cap Y|/|X * Y|.$$

It is zero if the sets are disjoint and one if they are the same. In practice, we use a lax threshold on Jaccard's coefficient such as 0.5 and ensure that all similarities have a p-value significance of at least 0.001. Specifically, we use the hypergeometric distribution to assess the likelihood of observing a given Jaccard's threshold (given the sizes of X and Y) and use this probability to derive a p-value test.

Given a database schema and two entity sets participating in it, e.g., "TFs" and "phenotypes", we first identify the paths between these entity sets in the underlying E/R diagram of the schema. Observe that there can be many paths, including recursive ones (e.g., "TFs regulate TFs which regulate other genes, contributing to phenotypes, when knocked-down."). Corresponding to each path, we instantiate a sequence of biclusterings and use the cover-tree to identify redescriptions that can link them into chains.

CMGSDB data sources and methods

We refer to the biological entities captured in CMGSDB as biots. CMGSDB contains exhaustive data about the following biots in *C. elegans*: chromosomes, genes, transcripts, and proteins. For genes, extensive annotations (IDs, locations, names, annotations, locus, transcripts) are complemented by microarray data, RNAi knockout experimental data, interaction data, gene regulatory information, and functional categorization

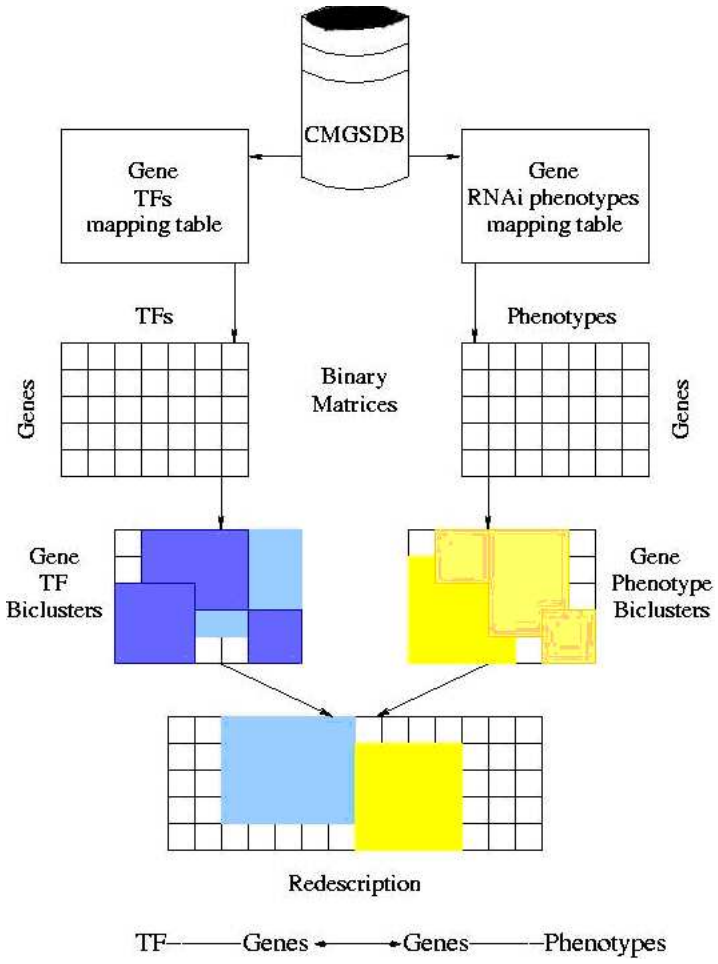


Figure A.1: Finding transcription factors (TFs) whose knock down induces improved desiccation tolerance in *C. elegans*. Two biclusters (shaded rectangles) joined at the gene interface using a redescription between their projections. Below that is the compositional data mining schema, displaying the sequence of primitives.

using the GO categories. Proteins, besides containing complete annotations, are enhanced by the addition of SwissProt/TrEMBL cross-references, physical structure details and properties, and orthology/paralogy information. Finally, groups of all types of biots and biot information are linked together by patterns found by compositional data mining, as described in the previous section.

Data Sources

Genome annotation data (chromosomes, genes, proteins, sequences, transcripts) for *C. elegans* are retrieved from Wormbase [43]. Attention has been paid to retaining all transcripts and their respective constituting coding sequences for each gene. These transcripts serve as a link to gene expression data and RNAi transcript information. Gene orthology and paralogy data have also been taken from Wormbase.

Protein sequences and annotations have been obtained from Wormbase, while their physical properties and PDB (Protein Data Bank [13]) homologs have been obtained from the SGCE (Structural Genomics of *C. elegans* [121]) project. Protein interaction data and gene regulatory information have been obtained from BioGRID [118]. Internal mappings from BioGRID IDs to Wormbase IDs have been generated.

Genomewide gene expression data for 496 *C. elegans* microarray experiments have been collected from SMD (Stanford Microarray Database [8]). Expression values have been related to the genes through gene transcripts.

The RNAi component of CMGSDB is one of the chief characteristics that discriminates CMGSDB from other *C. elegans* resources. The RNAi experiments obtained from Wormbase have been supplemented by RNAi experiments retrieved from Phenobank [117], PhenomicDB [47], and RNAi phenome database [83]. The same has been done for RNAi phenotypes. All RNAi phenotypes, thus obtained, have been organized into a hierarchical structure, with Body, Cell, Development, Lethal and Sterile, and Miscellaneous as the top phenotypic categories. While Phenobank's experiments test all *C. elegans* genes for their role in the first two rounds of mitotic cell division, RNAi phenome database's experiments are aimed at evaluating the effects of RNAi on genes whose knockdown causes embryonic lethality. PhenomicDB is a multi-organism phenotype-genotype database including human, mouse, fruit fly, *C. elegans*, and other model organisms. Apart from these web-based RNAi data sources, there are a number of genome-wide RNAi screens in literature that are undocumented in these web-based sources but have been included in CMGSDB ([21, 23, 26, 41, 45, 49, 51, 74, 76, 92, 98, 100, 105, 109, 111, 119, 120, 125, 128]).

Database schema

The key components of CMGSDB are illustrated in Figure A.2. Biots are contained in light green boxes, which are represented by one or more relations in CMGSDB. Blue arrows represent relationships in CMGSDB. Plain black arrows represent data flow.

Applying CDM to CMGSDB

We applied compositional data mining (CDM) to CMGSDB as follows. There are a variety of biclustering algorithms that can be applied for mining relationships [82]. For the purpose of this study, we utilized CHARM [133] to mine biclusters in binary relationships. For gene expression data, we utilized SAMBA [122] to mine biclusters.

Given a binary $0-1$ matrix, the CHARM algorithm identifies sets of rows that show the same bit (0/1) patterns across all columns. The row set is grown to be maximal in size and, together with the columns for which the rows have a "1", defines the bicluster. CHARM identifies overlapping biclusters, which can be organized alongside a lattice of subset relationships.

The SAMBA algorithm casts biclustering as a problem of finding bicliques in a bipartite graph. Given an edge-weighted graph (e.g., between genes and experiments labeled with expression levels) SAMBA detects dense subgraphs, which are then iteratively improved (using local addition/removal of vertices) in a post-processing phase.

Biclusters are connected if the overlap between the participating entities satisfied a Jaccard's threshold of 0.5. Chains computed in this manner all mediate through the Gene entity set, since it serves a central role in CMGSDB (i.e., all relationships involve Genes).

Patterns mined by CDM serve many purposes. For instance, they can be used to impute functions and properties to unannotated genes, they can make unexpected connections between upstream and downstream indicators, and they can summarize the distribution of data in the database more succinctly by identifying the sets of entities that dominate in many compositions.

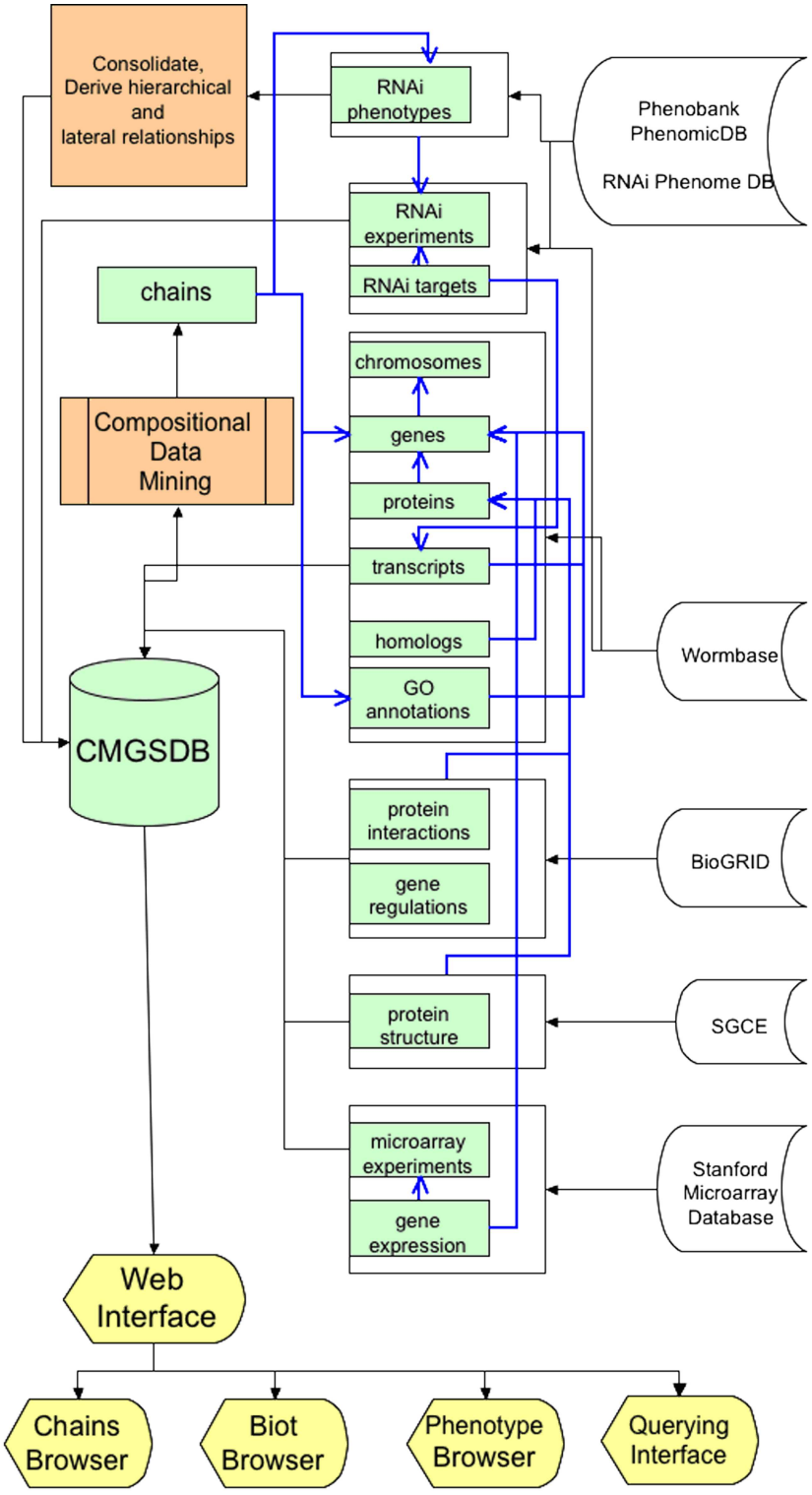


Figure A.2: Data integration and analysis in CMGSDB.

Table A.1: Summary of chain 153 containing gene *glp-1*

Bicluster	**Type**	**Set 1**	**Set 2**
1	Gene-Phenotype	*nmy-1*, *par-1*	PBPhen25 (Asymmetry of divison), WBPhen30 (Embryonic lethal), WBPhen301 (Protruding vulva), WBPhen320 (Sterile), WBPhen326 (Sterile progeny), WBPhen7 (Asymmetry of divison abnormal)
2	Gene-GO	*apx-1*, *glp-1*, *nmy-1*, *par-1*	GO:0002119 (Larval dev. (sensu Nematoda)), GO:0044464 (Cell part), GO:0009987 (Cellular process), GO:0048856 (Anatomical structure dev.), GO:0007389 (Pattern specification process) , GO:0009790 (Embryonic dev.), GO:0009791 (Post-embryonic dev.)
3	Gene-Gene	*glp-1*, *par-1*	*glp-1*, *par-1*

Querying CMGSDB

CMGSDB consists of a web interface and a PostgreSQL database management system. The web interface has been implemented using static and dynamic HTML, PHP, CSS, and Javascript. PostgreSQL is used to store the data described in the previous section and in Figure A.2.

The web interface of CMGSDB can be used for querying. The user can search against all *C. elegans* biots. Genes, for example, can be searched using names, loci, transcript IDs, and annotations. A biot page, apart from displaying basic information about that biot, also displays relationships with other biots that have been captured within CMGSDB. For instance, the phenotype page not only displays phenotype description, ID, and source, but also shows existing relationships with other phenotypes, GO categories associated with the phenotype, RNAi experiments in which the phenotype was observed, genes whose knock down resulted in the phenotype, and chains in which the phenotype participates. Biot pages are closely interlinked through biot IDs. As far as possible, biots are hyperlinked on pages. A biot page also contains hyperlinks to Wormbase and GO wherever applicable. Figure A.3 illustrates the page for the *gpr-1* gene through a screen shot.

Chains, as described before, are available for searching and browsing. Chains can be queried by participating genes, number of common genes among all biclusters, and number of biclusters. A chain with 3 biclusters containing gene *glp-1* is shown in Table A.1.

Details of *C. elegans* gene WBGene00001688

Wormbase ID:	WBGene00001688
Locus:	gpr-1
CDS Name:	F22B7.13
Transcript Name:	F22B7.13
Annotation:	gpr-1 encodes an extremely similar (97% identity) paralog of GPR-2; GPR-1 (and GPR-2) proteins have two N-terminal tetratricopeptide-like motifs and a C-terminal GoLoco/GPR (G protein regulatory) moti f, the latter of which has also been found in mamm alian AGS3 and Drosophila Pins; GPR-1 is required for normally asymmetrical cleavage of one-cell emb ryos; GPR-1 and GPR-2 form a high molecular weight (~700 kDa) complex that includes LIN-5; GPR-1 bin ds GDP-bound GOA-1 via a GoLoco/GPR motif, and dep ends on RIC-8 for this binding; GPR-1/2, GOA-1, an d LIN-5 colocalize at the cortex of early embryos; cortical enrichment of GPR-1 requires LIN-5, PAR- 2, PAR-3, and LET-99; the asymmetric distribution of GPR-1/2 and LET-99 in EMS cells is dependent on MES-1/SRC-1 signaling; GPR-1/2 is co-immunoprecip itated with RIC-8 and GPA-16.
Chromosome:	III
Starting Position:	8628838
Ending Position:	8630680
Strand:	1

Proteins associated with *C. elegans* gene WBGene00001688

CE24910

Transcripts associated with *C. elegans* gene WBGene00001688

Chromosome	Strand	Transcript Name	Exon	Coding Start Position	Coding End Position	Exon Start Position	Exon End Position
III	1	F22B7.13	exon1	8628860	8629107	8628838	8629107
III	1	F22B7.13	exon2	8629159	8629659	8629159	8629659
III	1	F22B7.13	exon3	8629710	8630317	8629710	8630317
III	1	F22B7.13	exon4	8630370	8630590	8630370	8630680

Protein-Protein interactions associated with *C. elegans* gene WBGene00001688

Gene/Protein A	Gene/Protein B	Direction	Experiment System	Pubmed IDs
WBGene00001648	WBGene00001688	AB	Two Hybrid	14704431
WBGene00001688	WBGene00017166	AB	Two Hybrid	14704431
WBGene00001688	WBGene00001686	AB	Two Hybrid	14704431
WBGene00001688	WBGene00002994	AB	Two Hybrid	14704431
WBGene00001688	WBGene00000754	AB	Two Hybrid	14704431
WBGene00000228	WBGene00001688	AB	Two Hybrid	14704431

Gene regulations associated with *C. elegans* gene WBGene00001688

Regulator Gene	Regulated gene	Pubmed IDs
WBGene00002994	WBGene00001688	10629219
WBGene00002994	WBGene00001688	8187641
WBGene00002994	WBGene00001688	631425
WBGene00002994	WBGene00001688	560330
WBGene00002994	WBGene00001688	7262539
WBGene00002994	WBGene00001688	7014288
WBGene00002994	WBGene00001688	7088142
WBGene00002994	WBGene00001688	6586368
WBGene00002994	WBGene00001688	6500256
WBGene00002994	WBGene00001688	2578115
WBGene00002994	WBGene00001688	1971988
WBGene00002994	WBGene00001688	2060028
WBGene00002994	WBGene00001688	10822257
WBGene00002994	WBGene00001688	12730122
WBGene00002994	WBGene00001688	12814548
WBGene00002994	WBGene00001688	14616061
WBGene00002994	WBGene00001688	15138888
WBGene00002994	WBGene00001688	11782949
WBGene00002994	WBGene00001688	12928525
WBGene00002994	WBGene00001688	1363076

RNAi Experiments associated with *C. elegans* gene WBGene00001688

PBRNAi1503705 WBRNAi00008207WBRNAi00024776WBRNAi00029651WBRNAi0003125
WBRNAi00042134WBRNAi00045289

RNAi Phenotypes associated with *C. elegans* gene WBGene00001688

PBPhen25	PBPhen28	WBPhen209	WBPhen30	WBPhen320
WBPhen329	WBPhen332	WBPhen48	WBPhen7	

Chains with *C. elegans* gene WBGene00001688

69	85	86	87	88
89	90	91	92	93
94	95	96	109	110
111	112	113	114	115
116	117	118	119	120
121	122	123	124	125

Figure A.3: Screenshot of the gene page.

LIN-12/Notch signaling

In *C. elegans*, the LIN-12/Notch protein family mediates cell-cell interactions. *Glp-1* and *lin-12* encode two proteins in the LIN-12/Notch pathway, which is conserved in mammalian development. The two general cell-cell interactions that determine cell-fate and involve these proteins are lateral specification and induction. Querying CMGSDB for *glp-1* gives two chains (chain 153 and chain 154). Table A.1 illustrates chain 153, which demonstrates a chain of 3 (2 non-trivial) biclusters. The biclusters with the GO categories and RNAi phenotypes suggest that genes in this chain contribute to the structural aspects of cell division such as pattern specification leading to asymmetry of division, and these might be important to avoid embryonic lethality, protruding vulva, and sterile progeny. Furthermore, this set of genes is likely to be self-regulated.

Four genes characterize the two chains: *par-1, apx-1, nmy-2, and glp-1. Par-1* encodes a serine threonine kinase, which is required for the spatial regulation of GLP-1 asymmetry [24]. *Par-1* is connected to *glp-1* through the GO and gene regulation blocks. *Apx-1* encodes a ligand homolog to the Delta protein of Drosophila. Both proteins contribute to the establishment of the dorsal-ventral axis in the early *C. elegans* embryo [6]. Chains 153 and 154 suggest an interaction between *par-1* and *apx-1*. The likelihood of this prediction is further strengthened by the computational prediction of interaction between the same pair of genes (or their products) by Zhong and Sternberg [136]. A putative gene in the Notch pathway is *nmy-2*, which encodes a maternally expressed non-muscle myosin II. The corresponding protein is linked through the phenotype bicluster containing *par-1*. The function of NMY-2 and PAR-5 is to together establish polarization in the *C. elegans* zygote along the anterior-posterior axis 23. In summary, *glp-1* and *par-1* interaction was already suggested, while *apx-1* and *nmy-2* represent new potential interactions with *glp-1* in the LIN-12/Notch pathway, uncovered through compositional data mining.

Wnt pathway

The Wnt signal transduction pathway regulates diverse processes including cell proliferation, migration, polarity, differentiation, and axon outgrowth in *C. elegans*. The signaling is composed of two pathways, the canonical wnt/BAR-1 pathway and the non-canonical wnt/WRM-1 pathway. A common component in both pathways is the HMG box containing protein POP-1, which is a member of the TCF/LEF family of transcription factors. The *wnt*-signaling pathway regulates the activation of the latter [78, 112]. CMGSDB reported 32 chains containing pop-1, the common target of the two wnt-pathways. These 32 chains suggested 18 new gene candidates (*daf-2, par-2, par-3, par-5, par-6, pkc-3, pkc-6, ooc-3, gpa-16, mbk-2, mes-1, csn-3, pgl-1, egl-46, tac-1, rab-5, tba-2, uri-1*) for the pathway. Of these, only par-5 (chains 234, 236, 240)

has been confirmed as a regulator of *pop-1* [80]. *pop-1* is connected to *par-2* (chains 204, 206, 210, 212) through a regulatory network [88, 131]. Consistent with the results from CMGSDB, Zhong and Sternberg [136] predicted interactions among *par-2*, *mes-1* (chains 246, 248), a gene encoding a tyrosine kinase-like protein that is required for unequal cell division [12], *ooc-3* (chains 222, 224), encoding a protein required to establish asymmetrical anterior-posterior cortical domains and spindle orientation [9], and *gpa-16* (chains 234, 236), encoding a member of the G-protein alpha-subunit family of heterochromatic GTPase that effects spindle position and orientation [1]. It can be hypothesized that PAR-2 is regulated by POP-1 over PAR-5. Further evidence shows that PAR-2 is regulated independently from the wnt-pathway, as it is not regulated by MOM-5 and MOM-2, the wnt-receptor and wnt-ligand respectively [75]. From the above gene list of 18 genes, CMGSDB suggests an interaction of wnt-proteins with the tyrosine kinase receptor DAF-2, which is involved in longevity and insulin signaling. This can be a potential link between daf-proteins and wnt-pathway proteins, indicating a possible connection between insulin and wnt signaling.

Some database statistics

In this section, we describe some basic statistics about the data in CMGSDB, especially focusing on data related to RNAi experiments and phenotypes and chains. Figure A.4 illustrates some of the statistics of chains. Chains consisting of 3, 4, and 5 biclusters, number 2054, 1654, and 426, respectively. Figure A.4 examines the distribution of the total number of genes in a chain and the number of common genes among all biclusters in a chain.

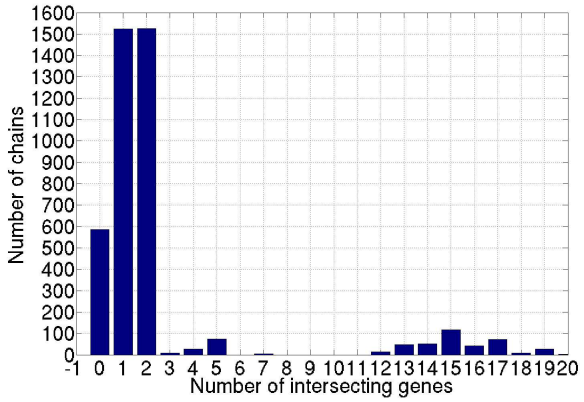
CMGSDB stores 81722 RNAi experiments and 565 RNAi phenotypes. This includes 145028 relationships between 21222 unique *C. elegans* gene transcripts and the above 565 phenotypes.

Phenotype browser

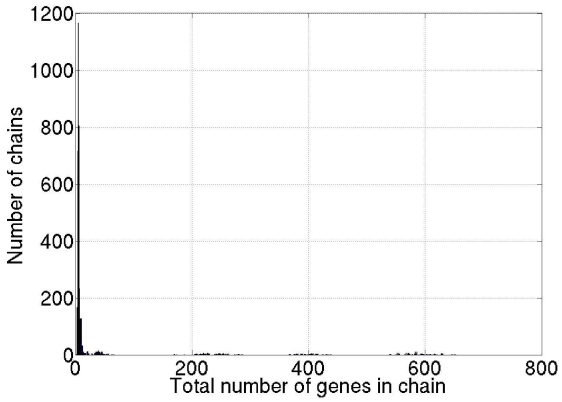
In CMGSDB, phenotypes from several different sources have been organized into a common hierarchy. This hierarchy is available for browsing via a phenotype browser available at

`https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/Treeview/index.php`.

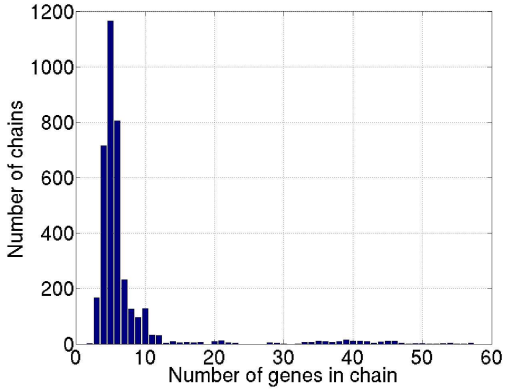
The viewer has been implemented using the PHP TreeView class and is dynamically linked to individual phenotype pages and to other biots. Figure A.5 illustrates the phenotype browser with the tree view on the left.



(a)



(b)



(c)

Figure A.4: Statistics of chains. (a) Distribution of number of common genes in a chain. (b) Distribution of total number of genes in a chain. (c) Subset of (b).

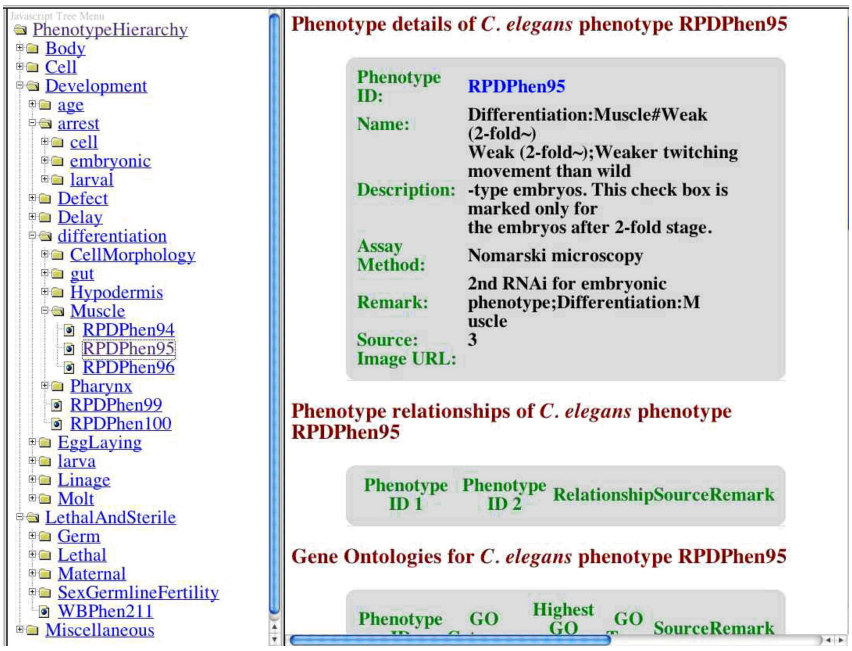


Figure A.5: Screenshot of the phenotype browser.

Downloads

We have made the CMGSDB schema, scripts, and raw data freely available under the GPL. Only the software for computing chains is not included. The download package is available at

`https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/download.php`.

Using this package, a user with proper hardware and software resources (including PostgreSQL and Perl) can locally set up an exact replica of CMGSDB's back end. The data is downloaded at runtime dynamically over the Internet. Scripts prepare the data and populate the database. This includes the integration of phenotypes from various sources.

All data in CMGSDB (except data related to chains) is available for download as flat files at the downloads page.

Concluding remarks

The integration of RNAi data and the application of data mining within CMSGDB provide the user with enhanced abilities to interpret raw *C. elegans* data. Unlike existing *C. elegans* resources, CMGSDB integrates RNAi data from multiple discrete sources. Using chains, users can discover new associations and relationships in the data that can be tested experimentally. A very meaningful future direction is to further consolidate the phenotypes to support alternate sets of phenotypes. This could be done by identifying very similar phenotypes as the same or by choosing a level of specialization in the phenotype tree. During the final two years of the CMGS project, additional data mining and modeling capabilities will be added.