Map-Invariant Spectral Analysis for the Identification of DNA Periodicities

Ahmad Rushdi*[1] and Jamal Tuqan[2]and Thomas Strohmer[3]

[1]Author was with the department of Electrical and Computer Engineering at the University of California, Davis, CA 95616, USA, and is now with Cisco Systems, Inc. San Jose CA 95134, USA.
[2]Author was with the department of Electrical and Computer Engineering at the University of California, Davis, CA 95616, USA.
[3]Author is with the department of Mathematics, University of California, Davis, CA 95616, USA. T.S. acknowledges partial support from the NSF via grants DMS 0811169 and DMS-1042939.

Email: Ahmad Rushdi*- aarushdi@ieee.org;

*Corresponding author

Abstract

Many signal processing based methods for finding hidden periodicities in DNA sequences have primarily focused on assigning numerical values to the symbolic DNA sequence and then applying spectral analysis tools such as the short-time discrete Fourier transform (ST-DFT) to locate these repeats. The key results pertaining to this approach are however obtained using a very specific symbolic to numerical map, namely the so-called Voss representation. An important research problem is to therefore quantify the sensitivity of these results to the choice of the symbolic to numerical map. In this paper, a novel algebraic approach to the periodicity detection problem is presented and provides a natural framework for studying the role of the symbolic to numerical map in finding these repeats. More specifically, we derive a new matrix-based expression of the DNA spectrum that comprises most of the widely used mappings in the literature as special cases, shows that the DNA spectrum is in fact invariable under all these mappings, and generates a necessary and sufficient condition for the invariance of the DNA spectrum to the symbolic to numerical map. Furthermore, the new algebraic framework decomposes the periodicity detection problem into several fundamental building blocks that are totally independent of each other. Sophisticated digital filters and/or alternate fast data transforms such as the discrete cosine and sine transforms can therefore be always incorporated in the periodicity detection scheme regardless of the choice of the symbolic to numerical map. Although the newly proposed framework is matrix based, identification of these periodicities can be achieved at a low computational cost.

1 Introduction

Many researchers have noted that the occurrence of repetitive structures in a DNA sequence is symptomatic of a biological phenomena. Specific applications of this observation include identification of diseases [1], DNA forensics [2], and detection of pathogen exposure [3]. Some of these structures are simple repetition of short DNA segments such as exons [4], tandem repeats [5], dispersed repeats [6], and unstable triplet repeats in the noncoding regions [7] while other forms more elaborate patterns such as palindromes [8] and the period-3 component [9–13], a strong periodic characteristic found primarily in genes and pseudogenes [14]. Methods that detect these DNA periodicities are either probabilistic or deterministic. Most of the deterministic techniques rely on spectral analysis of the DNA sequence using the short-time discrete Fourier transform (ST-DFT) [15–17]. The main idea is as follows: given a DNA sequence of length N, numerical values are first assigned to every element in $\mathbb{F} = \{A, C, G, T\}$ where these letters denote the four nucleotides in the DNA, namely the two purines: adenine (A) and guanine (G) and the two pyrimidines: thymine (T) and cytosine (C). A typical DNA double helix is shown in Figure 1. The symbolic to numerical map is clearly not unique,

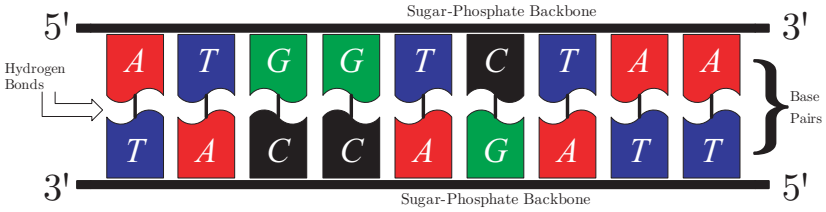


Figure 1: DNA: a straightened helix structure.

typically has a biological interpretation, and needs to preserve the specific structure of the DNA sequence under study. One such popular map is the Voss representation $\mathbb{F} \longmapsto \mathbb{D} = \{0, 1\}$ where four binary indicator sequences $x_l(n)$, $l \in \mathbb{F}$, are generated with 1 indicating the presence of a nucleotide and 0 its absence [18]. An example of the mapping of a single DNA strand to $x_l(n), \forall\ l \in \mathbb{F}$ is shown in Figure 2. Once the DNA symbolic sequence is mapped into numerical version(s), a set of discrete time sequences are generated and are the numerical equivalence of the DNA sequence. These numerical sequences can then by processed using standard signal processing techniques. In particular, the ST-DFT for each elementary sequences can be computed as

$$X_l(Rn, k) \triangleq \sum_{m=-M+1}^{0} x_l(Rn + m)h(m)e^{-j2\pi mk/M}, \tag{1}$$

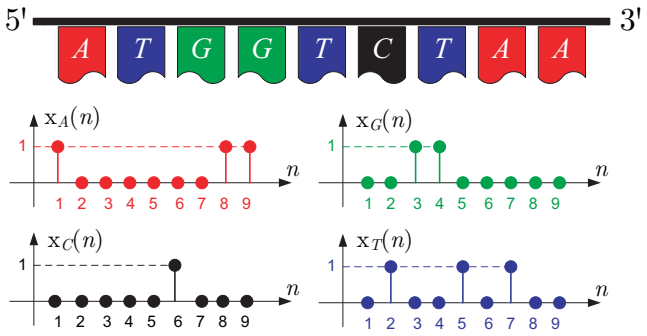


Figure 2: The Voss representation of a DNA segment.

where n is the window starting point, R is the amount of window shift, and $h(m) = 1$ for $-M+1 \leq m \leq 0$ and zero otherwise. If $R = 1$, then, the window slides one nucleotide at a time whereas if $R = 3$, the displacement of the window is on a 3-nucleotide basis. Note that the all-ones function $h(m)$ does not affect the value of $X_l(Rn, k)$. However, it serves as a place holder for other filters that can be used to replace it, as will be shown in the next section. One popular application of the ST-DFT based technique that has received considerable attention in the past is the identification of the period-3 component using the DNA spectrum, defined for $R = 3$ as follows

$$S(n) = \sum_{l \in \mathbb{F}} |X_l(3n, \frac{M}{3})|^2 = \sum_{l \in \mathbb{F}} |\sum_{m=-M+1}^{0} x_l(3n+m) e^{-j2\pi m/3}|^2. \tag{2}$$

A number of researchers have advocated the use of the period-3 component to discriminate between coding and non coding regions (see for example [11, 13, 16, 19–23] to name a few) but the subject remains highly controversial as it is successful for certain genes but does not work for others. To better comprehend the underlying reasons behind this disparity in performance, a new multirate DSP model that provides a full understanding of the inner workings of the DNA periodicity has been first proposed in [24], and studied in details in [25]. This model is shown in Figure 3. This model provides closed form expressions for the DNA spectrum that generalize and unify some of the already existing results in the literature were obtained. One of these expressions in particular clearly shows that the identification of the period-3 component in the DNA spectrum, a signal processing problem, is equivalent to the detection of the nucleotide distribution disparity in the codon structure of a DNA sequence, a genomic problem. The disparity in the nucleotide distribution within the codon structure of a DNA sequence is termed the codon bias. Using this model, the DNA spectrum is completely characterized by a set of digital sequences, termed the *filtered polyphase sequences*. By processing these sequences, signal processing techniques can potentially have an impact on understanding and detecting biological structures of this nature. From a computational cost perspective,

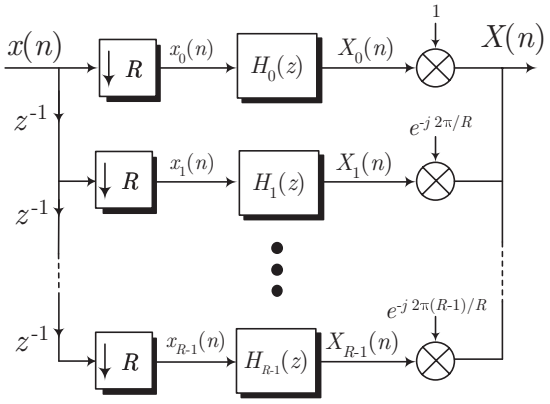


Figure 3: The Multirate DSP model for general R. The period-3 case is easily obtained by setting $R = 3$.

the computation of the DNA spectrum using this model does not require any complex valued operations [26]. This finding is rather surprising given the existence of complex multipliers in the proposed DSP model as clearly illustrated in Figure 3. It is shown that the direct computation of the DNA spectrum using (2) requires essentially double the amount of arithmetic operations compared to the DSP model approach.

It is important, however, to keep in mind that the above conclusions and results were obtained using the Voss symbolic to numerical transformation. A fundamental research issue is to therefore determine the sensitivity of the signal processing based method to the choice of the *symbolic to numerical map*. In particular, the core question here is: how dependent are the above results on the Voss representation? Are these results invariant with respect to the other popular maps in the literature? Can we derive necessary and/or sufficient conditions for the invariance of the DNA spectrum to the symbolic to numerical transformation? Is there a general mathematical framework that can help us generate new symbolic to numerical maps for which the DNA spectrum remains essentially the same? These are the type of questions we address in this paper and provide answers to. One approach to answer this question was presented in [27], where a novel framework for the analysis of the equivalence of the mappings used for numerical representation of symbolic data based on signal correlation was presented, along with strong and weak equivalence properties. In [28], we attempted to answer the same question starting at the aforementioned DSP model for a limited set of mappings. Our main goal in this work is to deembed the symbolic to numerical mapping process from the DNA spectrum computation process. We answer a set of other relevant questions along the way.

A key remark is in order at this point: while the DSP model approach proposed in Figure 3 has many

advantages, it is not well suited for investigating the role of the symbolic to numerical map in the identification of DNA harmonics. It follows that *a completely new paradigm for detecting DNA harmonics is required*. The main contribution of this paper is therefore the derivation of a novel matrix-based framework for the computation of the DNA spectrum that is extremely well fitted to the study of the symbolic to numerical transformation. Specifically, we first derive a new matrix-based expression of the DNA spectrum that:

1. comprises most of the existing mappings in the literature as special cases,
2. shows that the DNA spectrum is in fact invariable under all these mappings,
3. generates a necessary condition for the invariance of the DNA spectrum to the symbolic to numerical mapping used to compute it.

Furthermore, the new algebraic framework presented here decomposes the frequency identification problem into several fundamental components that are *totally independent of each other*. It follows that sophisticated digital filters and/or alternative transformations to the DFT such as the discrete cosine, sine, and Hartley transforms can *always* be easily incorporated in the harmonics detection scheme irrespective of the choice of the symbolic to numerical map. Finally, although the newly proposed framework is matrix based, we show that similar to the DSP model approach, the computation of the DNA spectrum using this new framework is very efficient.

The paper is organized as follows. In section 2, we derive a new matrix based framework to efficiently compute the ST-DFT-based spectrum. New expressions for the ST-DFT $X_l(Rn, \frac{M}{R})$ and its magnitude squared $|X_l(Rn, \frac{M}{R})|^2$ are obtained and indicate that these quantities are completely parameterized by some pre-defined matrices. The numerical values of these matrices simply depend on our choice of filtering (e.g. rectangular window versus non-rectangular one versus general FIR filters) as well as our choice of data transform (e.g. the DFT versus the DCT versus the DST). Using these results, in section 3, a new expression of the DNA power spectrum is derived and is also completely defined by these matrices. The elegance of this matrix based approach is that it allows the incorporation of general symbolic to numerical maps into the newly derived DNA spectrum expression *provided these generic maps can be expressed as affine transformations of the Voss representation*. This last assumption is motivated by the fact that all the popular maps that are available in the literature satisfy the affine condition. Furthermore, the maps are now completely characterized by the affine transformation (two matrices $\mathbf{A}$ and $\mathbf{b}$) and can be therefore

changed *without affecting the remaining matrices* in the DNA spectrum expression. In conclusion, the newly derived DNA spectrum expression is stated as a function of a number of matrices. Each of these matrices captures an essential component of the process (filtering, data transform, symbolic to numerical map) and the elements of each matrix can be changed without affecting the other matrices. In section 4 and using the above results, we show that the Voss-based DNA spectrum is essentially invariant under some of the most popular maps in the literature. A **necessary and sufficient** condition for the invariance of the DNA spectrum under any map is also derived. In section 5, we show how the special structure of the filtering matrix allows the efficient use of sophisticated digital filters to improve the detection performance of DNA harmonics through the computation of the DNA spectrum. We also show how to replace the DFT by other fast transforms such as the discrete cosine transform (DCT), the discrete sine transform (DST), and the discrete Hartly transform (DHT). Finally, some concluding remarks are mentioned in section 6. A list of the different notation used in the paper is summarized in Table 1.

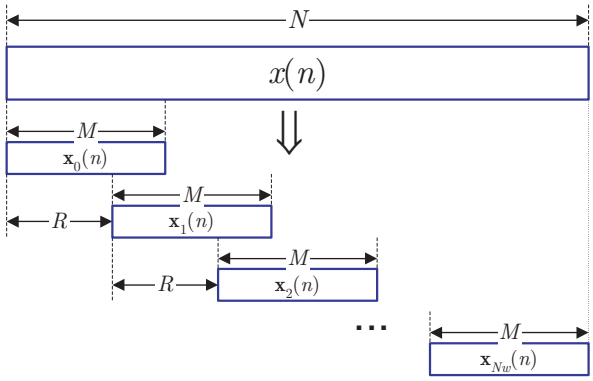


Figure 4: Splitting $x(n)$ into N_w overlapping sections $\mathbf{x}_r(n)$ using a sliding window approach.

2 A New algebraic framework for computing the ST-DFT

Given a sequence $x(n)$ of length N, the ST-DFT is typically implemented using a sliding window approach as shown in Figure 4. Windows of length M that overlap with a factor R are first generated to form $\mathbf{x}_r(n), r = 1, 2, \ldots, N_w$, where $N_w = \lceil (N - M + 1)/R \rceil$ is the number of resulting windows. Once we map the DNA sequence into an integer number of numeric sequences γ, given by $x_l(n),\ l = 1, \ldots, \gamma$ ($\mathbb{F} \longmapsto \mathbb{D}$), the ST-DFT's $X_l(n),\ l = 1, \ldots, \gamma$ can be found and their squared magnitudes are added to result in the DNA Spectrum $S(n)$ as summarized in Figure 5.

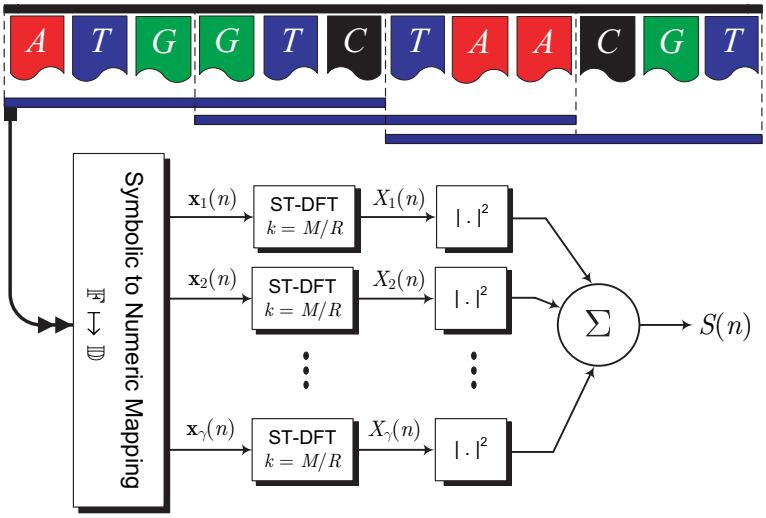


Figure 5: System structure to find the DNA power spectrum $S(n)$ by extracting successive sliding windows of the symbolic DNA sequence, mapping each to γ numeric sequences, finding their DFT's at $k = \frac{M}{R}$, and finally adding the corresponding squared magnitudes. In this example, $N_w = \lceil (N-M+1)/R \rceil = \lceil (12-6+1)/3 \rceil = 3$ windows are generated.

It was shown in [26] that the ST-DFT of $x(n)$ can be written as

$$X(Rn, \frac{M}{R}) = X_0(n) + X_1(n)e^{-j\frac{2\pi}{R}} + \cdots + X_{R-1}(n)e^{-j2\pi\frac{R-1}{R}}, \tag{3}$$

where the quantities $X_r(n), \forall\ r \in \{0, 1, \ldots, R-1\}$ are the so-called filtered polyphase sequences given by

$$X_r(n) \doteq X_r(Rn, \frac{M}{R}) = \sum_{m=r, r+R, \ldots}^{\lfloor \frac{M}{R}-1 \rfloor} x(Rn + Rm + r)h_r(m), \tag{4}$$

$\forall\ r \in \{0, 1, \ldots, R-1\}$. The impulse response $h_r(m)$ is the inverse $\mathcal{Z}$-transform of $H_r(z)$ in Figure 3. Equations (3) and (4) can be used to compute the ST-DFT of a discrete time sequence, and subsequently its magnitude squared. In this section, we re-express these equations in matrix form, and then use the new formula to derive an expression for $|X(Rn, \frac{M}{R})|^2$. Throughout the paper, vectors and matrices (arrays) are always expressed in bold letters. The notation for the various matrix operations is given in Table 2.

2.1 Matrix formulation of the ST-DFT

Using the defined matrix notation, we can restate equation (3) as

$$X(Rn, \frac{M}{R}) = \begin{bmatrix} 1 & e^{-j\frac{2\pi}{R}} & \ldots & e^{-j2\pi\frac{R-1}{R}} \end{bmatrix} \begin{bmatrix} X_0(n) \\ X_1(n) \\ \vdots \\ X_{R-1}(n) \end{bmatrix} \doteq \mathbf{C}^T \mathbf{\Gamma}(n). \tag{5}$$

The real valued array

$$\mathbf{\Gamma}(n) \quad = \quad [X_0(n) \quad X_1(n) \quad \ldots \quad X_{R-1}(n)]^T \tag{6}$$

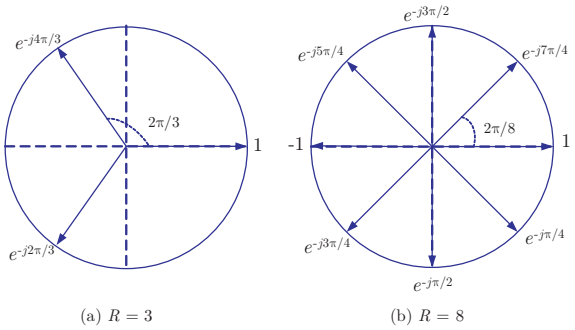
is the vector whose elements are the R filtered polyphase components. Similarly, the complex valued R-element array

$$\mathbf{C} \quad = \quad \left[\begin{array}{cccc} 1 & e^{-j\frac{2\pi}{R}} & \ldots & e^{-j2\pi\frac{R-1}{R}} \end{array} \right]^T \tag{7}$$

is the vector whose elements are the R equispaced phasors located on the unit circle with $\frac{2\pi}{R}$ phase deviations as shown in Figure 6 for $R = 3$ and $R = 8$. Note that

$$\sum_{r=0}^{R-1} e^{-j2\pi r/R} = \frac{1 - (e^{-j2\pi/R})^R}{1 - e^{-j2\pi/R}} = 0, \tag{8}$$

$\forall\ R \neq 1$, which implies that the sum of elements in $\mathbf{C}$ is equal to 0. This is a key feature of the complex array $\mathbf{C}$ that will be used in later sections to simplify important expressions. On the other hand, we observe



(a) $R = 3$ (b) $R = 8$

Figure 6: Elements of array $\mathbf{C}$ of Equation (7), represented as phasors on the unit circle for (a) $R = 3$, and (b) $R = 8$.

that (4) can be written in the following matrix format

$$X_r(n) = \left[\begin{array}{cccc} 1 & 1 & \ldots & 1 \end{array} \right] \left[\begin{array}{c} x(Rn + r) \\ x(Rn + r + R) \\ \vdots \\ x(Rn + r + M - R) \end{array} \right] \doteq \mathbf{h}^T \hat{\mathbf{x}}_r(n), \tag{9}$$

$\forall\ r \in \{0, 1, \ldots, R-1\}$, where $\mathbf{h}$ is an all-one vector of length M/R, and $\hat{\mathbf{x}}_r(n)$ of length M/R is the r^{th} polyphase component of the window $\mathbf{x}(n)$ of length M. Using (9), the R filtered polyphase components $X_r(n)$ can be arranged in the following array format

$$[X_0(n) \quad X_1(n) \quad \ldots \quad X_{R-1}(n)] = \mathbf{h}^T [\hat{\mathbf{x}}_0(n) \quad \hat{\mathbf{x}}_1(n) \quad \ldots \quad \hat{\mathbf{x}}_{R-1}(n)]. \tag{10}$$

Using the identity

$$vec(\mathbf{A}_1\mathbf{A}_2) = (\mathbf{I} \otimes \mathbf{A}_1)vec(\mathbf{A}_2), \tag{11}$$

it follows that

$$\begin{bmatrix} X_0(n) \\ X_1(n) \\ \vdots \\ X_{R-1}(n) \end{bmatrix} = \left(\mathbf{I}_R \otimes \mathbf{h}^T\right) \begin{bmatrix} \hat{\mathbf{x}}_0(n) \\ \hat{\mathbf{x}}_1(n) \\ \vdots \\ \hat{\mathbf{x}}_{R-1}(n) \end{bmatrix},$$

which can be restated in matrix format as

$$\mathbf{\Gamma}(n) = \left(\mathbf{I}_R \otimes \mathbf{h}^T\right) \hat{\mathbf{x}}(n) = \mathbf{H}\ \hat{\mathbf{x}}(n), \tag{12}$$

where $\mathbf{H} \doteq \mathbf{I}_R \otimes \mathbf{h}^T$ is an $R \times R$ matrix of $1 \times \frac{M}{R}$ blocks, given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}^T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{h}^T & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{h}^T \end{bmatrix} = \begin{bmatrix} 1\ 1\ ..\ 1 & \dots & 0\ 0\ ..\ 0 \\ \vdots & \ddots & \vdots \\ 0\ 0\ ..\ 0 & \dots & \underbrace{1\ 1\ ..\ 1}_{M/R} \end{bmatrix}.$$

The window $\hat{\mathbf{x}}(n)$ of length M is a block interleaved version of the sliding window $\mathbf{x}(n)$ of length M starting at index n. Generating $\hat{\mathbf{x}}(n)$ can be accomplished by blocking the window $\mathbf{x}(n)$ into an array of R elements per row (hence M/R rows), and then reading the array out column by column. The ST-DFT $X(Rn, \frac{M}{R})$ can therefore be completely identified as a function of $\mathbf{C}$, $\mathbf{h}$, and $\hat{\mathbf{x}}(n)$ as follows

$$X(Rn, \frac{M}{R}) = \mathbf{C}^T \left(\mathbf{I}_R \otimes \mathbf{h}^T\right) \hat{\mathbf{x}}(n) = \mathbf{C}^T\mathbf{H}\ \hat{\mathbf{x}}(n). \tag{13}$$

The complex row vector $\mathbf{C}^T\mathbf{H}$ is an array of R blocks, each of length $\frac{M}{R}$ as given by

$$\mathbf{C}^T\mathbf{H} = \left[\underbrace{1\ ..\ 1}_{M/R}\ \underbrace{e^{-j\frac{2\pi}{R}}\ ..\ e^{-j\frac{2\pi}{R}}}_{M/R}\ \dots\ \underbrace{e^{-j2\pi\frac{R-1}{R}}\ ..\ e^{-j2\pi\frac{R-1}{R}}}_{M/R}\right],$$

which represents M/R repetitions of the elements in $\mathbf{C}$. Similar to $\mathbf{C}$, the sum of elements in $\mathbf{C}^T\mathbf{H}$ is equal to 0.

2.2 A matrix based expression for the magnitude squared of the ST-DFT

Using (5), the magnitude squared of the ST-DFT can be expressed as

$$|X(Rn, \frac{M}{R})|^2 = X^H(n)X(n) \doteq \mathbf{\Gamma}^H(n)\mathbf{D}\mathbf{\Gamma}(n), \tag{14}$$

where matrix $\mathbf{D} \doteq \mathbf{C}^{\star}\mathbf{C}^{T}$ is an $R \times R$ matrix given by

$$\mathbf{D} = \begin{bmatrix} 1 & e^{-j\frac{2\pi}{R}} & \dots & e^{-j2\pi\frac{R-1}{R}} \\ e^{j\frac{2\pi}{R}} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-j\frac{2\pi}{R}} \\ e^{j2\pi\frac{R-1}{R}} & \dots & e^{j\frac{2\pi}{R}} & 1 \end{bmatrix}.$$

$\mathbf{D}$ is obviously a right circulant (hence Toeplitz) matrix whose rows and columns are rotated versions of $\mathbf{C}$. Obviously, the sum of any row or column elements in $\mathbf{D}$ is equal to 0. Substituting (12) in (14), or equivalently using (13), implies that the spectrum $S(n)$ can be stated as

$$\begin{aligned} |X(Rn, \frac{M}{R})|^2 &= \hat{\mathbf{x}}^H(n)[\left(\mathbf{I}_R \otimes \mathbf{h}^T\right)^H \mathbf{C}^{\star}][\mathbf{C}^T \left(\mathbf{I}_R \otimes \mathbf{h}^T\right)]\hat{\mathbf{x}}(n) \\ &= \hat{\mathbf{x}}^H(n)\ \mathbf{H}^H\mathbf{D}\ \mathbf{H}\ \hat{\mathbf{x}}(n) \\ &= \hat{\mathbf{x}}^H(n)\ \mathbf{W}\ \hat{\mathbf{x}}(n), \end{aligned} \tag{15}$$

where

$$\mathbf{W} \doteq \mathbf{H}^H\mathbf{D}\mathbf{H} = (\mathbf{C}^T\mathbf{H})^H(\mathbf{C}^T\mathbf{H}),$$

is an $R \times R$ matrix of $\frac{M}{R} \times \frac{M}{R}$ blocks, given by

$$\mathbf{W} = \begin{bmatrix} \mathbf{1} & \mathbf{e}^{-j\frac{2\pi}{R}} & \dots & \mathbf{e}^{-j2\pi\frac{R-1}{R}} \\ \mathbf{e}^{j\frac{2\pi}{R}} & \mathbf{1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{e}^{-j\frac{2\pi}{R}} \\ \underbrace{\mathbf{e}^{j2\pi\frac{R-1}{R}}}_{\frac{M}{R}\times\frac{M}{R}} & \dots & \mathbf{e}^{j\frac{2\pi}{R}} & \underbrace{\mathbf{1}}_{\frac{M}{R}\times\frac{M}{R}} \end{bmatrix}.$$

Matrix $\mathbf{W}$ can be represented as a Kronecker product of $\mathbf{D}$ and an $\frac{M}{R} \times \frac{M}{R}$ all-one matrix. Note that any row or column in $\mathbf{W}$ is a rotated version of $\mathbf{C}^T\mathbf{H}$, therefore, the sum of the elements of any row or column in $\mathbf{W}$ is equal to 0.

3 The New DNA Spectrum Expression

A first step towards finding the DNA spectrum $S(n)$ is the symbolic to numeric mapping $\mathbb{F} \longmapsto \mathbb{D}$ as was shown in Figure 5. Once the symbolic DNA sequence is mapped into γ numeric sequence(s), the short-time discrete Fourier transform is applied to each of them and the sum of the squared magnitudes of the ST-DFTs will result in the DNA spectrum at the frequency point $k = \frac{M}{R}$ as given by

$$S(Rn, k)|_{k=\frac{M}{R}} = \sum_{l=1}^{\gamma} |X_l(Rn, \frac{M}{R})|^2. \tag{16}$$

For simplicity, we denote $S(Rn,k)|_{k=\frac{M}{R}}$ as $S(n)$ in the next sections. Several mappings were introduced in the literature using both real and complex numerical values with typical number of sequences $\gamma = 1$ up to 4 to maintain reasonable computation complexity. In this section, we use the results of section 2 to derive general expressions for the M/R ST-DFT and spectrum for any symbolic to numeric mapping.

3.1 The Voss-based DNA Spectrum

The simplest and most commonly used map of a DNA sequence is the Voss representation $\mathbb{F} \longmapsto \mathbb{V}$: that is to form $\gamma = 4$ binary indicator sequences $x_A(n)$, $x_C(n)$, $x_G(n)$, and $x_T(n)$ where a 1 would indicate the presence of a base and 0 indicates its absence [18]. This approach has been extensively used in relevant genomic research. Note that the four sequences are not linearly independent since for any index n, the four sequences will add up to one. That is

$$x_A(n) + x_C(n) + x_G(n) + x_T(n) = 1.$$

This redundancy plays an important role in the derivations of this section. Moreover, it follows that for any length-M window starting at n, the four mapped Voss windows will add up to an all-one length-M sequence and the same fact holds for the interleaved windows

$$\begin{aligned} \mathbf{x}_A(n) + \mathbf{x}_C(n) + \mathbf{x}_G(n) + \mathbf{x}_T(n) &= \hat{\mathbf{x}}_A(n) + \hat{\mathbf{x}}_C(n) + \hat{\mathbf{x}}_G(n) + \hat{\mathbf{x}}_T(n) \\ &= [1\ 1\ \ldots\ 1]^T. \end{aligned} \tag{17}$$

For illustration, Figure 7(a) shows a sample DNA window that is mapped into the corresponding numeric windows $\mathbf{x}_l(n), \forall\ l \in \mathbb{F}$ in Figures 7(b), 7(d), 7(f), and 7(h). With an example interleaving factor $R = 3$, the interleaved windows $\hat{\mathbf{x}}_l(n), \forall\ l \in \mathbb{F}$ are shown in Figures 7(c), 7(e), 7(g), and 7(i). Each of the four sequences is a discrete time sequence that can be processed using the analysis of section 2. Therefore, the ST-DFT of each sequence can be found using (13) to be

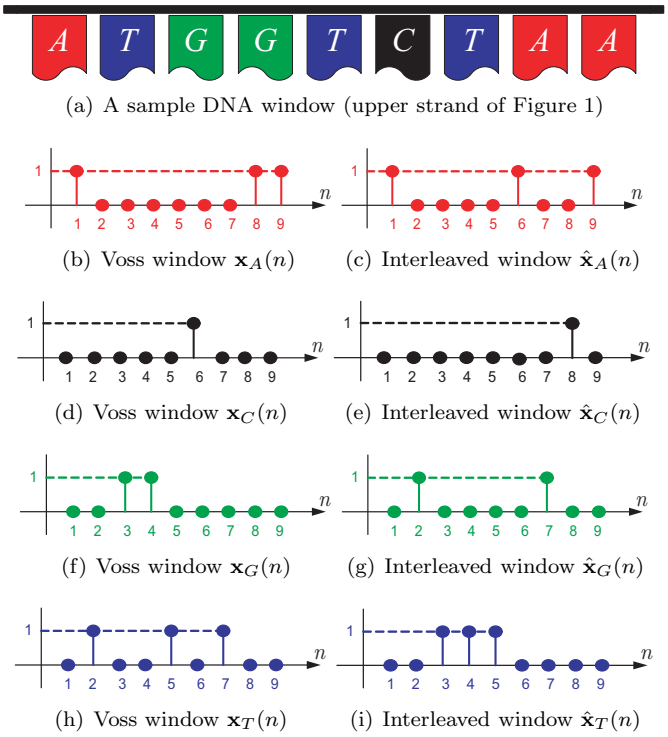
$$X_l(n) = \mathbf{C}^T\mathbf{H}\ \hat{\mathbf{x}}_l(n), \tag{18}$$

$\forall\ l \in \mathbb{F}$, and the power spectrum of each sequence can hence be derived as in (15) to be

$$S_l(n) = |X_l(n)|^2 = \hat{\mathbf{x}}_l^H(n)\ \mathbf{W}\ \hat{\mathbf{x}}_l(n),$$

$\forall\ l \in \mathbb{F}$. It follows that the Voss-based DNA spectrum $S_v(n)$ is

$$\begin{aligned} S_v(n) &\doteq |X_A(n)|^2 + |X_C(n)|^2 + |X_G(n)|^2 + |X_T(n)|^2 \\ &= \sum_{l\in\mathbb{F}} \hat{\mathbf{x}}_l^H(n)\ \mathbf{W}\ \hat{\mathbf{x}}_l(n). \end{aligned} \tag{19}$$



(a) A sample DNA window (upper strand of Figure 1)

(b) Voss window $\mathbf{x}_A(n)$ (c) Interleaved window $\hat{\mathbf{x}}_A(n)$

(d) Voss window $\mathbf{x}_C(n)$ (e) Interleaved window $\hat{\mathbf{x}}_C(n)$

(f) Voss window $\mathbf{x}_G(n)$ (g) Interleaved window $\hat{\mathbf{x}}_G(n)$

(h) Voss window $\mathbf{x}_T(n)$ (i) Interleaved window $\hat{\mathbf{x}}_T(n)$

Figure 7: A sample DNA window of length $M = 9$, the corresponding Voss binary windows $\mathbf{x}_l(n), \forall\ l \in \mathbb{F}$, and the interleaved versions $\hat{\mathbf{x}}_l(n), \forall\ l \in \mathbb{F}$ with an interleaving factor $R = 3$. The interleaved windows are generated by rearranging the original windows in an $R = 3$-interleaved format. In this example, data points of $\hat{\mathbf{x}}_l(n)$ at $(1, 2, 3), (4, 5, 6), (7, 8, 9)$ are mapped from those in $\mathbf{x}_l(n)$ at $(1, 4, 7), (2, 5, 8), (3, 6, 9)$.

An obvious step at this point is to simplify (19) to avoid the summation over different bases. To do this, we use Equation (18) to arrange the ST-DFT's of $x_l(n)$, $\forall\ l \in \mathbb{F}$ in the following format

$$[X_A(n)\ \ X_C(n)\ \ X_G(n)\ \ X_T(n)] = \mathbf{C}^T\mathbf{H}\ \ [\hat{\mathbf{x}}_A(n)\ \ \hat{\mathbf{x}}_C(n)\ \ \hat{\mathbf{x}}_G(n)\ \ \hat{\mathbf{x}}_T(n)]. \tag{20}$$

Using (11), it follows that

$$\begin{bmatrix} X_A(n) \\ X_C(n) \\ X_G(n) \\ X_T(n) \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \otimes \mathbf{C}^T\mathbf{H} \right) \cdot \begin{bmatrix} \hat{\mathbf{x}}_A(n) \\ \hat{\mathbf{x}}_C(n) \\ \hat{\mathbf{x}}_G(n) \\ \hat{\mathbf{x}}_T(n) \end{bmatrix}.$$

We define $\boldsymbol{\Upsilon}_v(n)$: the array of the four Voss-based ST-DFTs. It can now be written as

$$\boldsymbol{\Upsilon}_v(n) = \left[\ X_A(n) \quad X_C(n) \quad X_G(n) \quad X_T(n)\ \right]^T = \left(\mathbf{I}_4 \otimes \mathbf{C}^T\mathbf{H}\right) \hat{\mathbf{x}}_v(n), \tag{21}$$

where $\mathbf{I}_4$ is the 4×4 identity matrix, and the vector $\hat{\mathbf{x}}_v(n)$ of length $4M$ is an array of the four Voss interleaved windows starting at index n: $\hat{\mathbf{x}}_l(n)$, $\forall\ l \in \mathbb{F}$. Using the identity

$$(\mathbf{A}_1 \otimes \mathbf{A}_2)(\mathbf{A}_3 \otimes \mathbf{A}_4) = (\mathbf{A}_1\mathbf{A}_3 \otimes \mathbf{A}_2\mathbf{A}_4), \tag{22}$$

the Voss-based DNA power spectrum can be manipulated into

$$\begin{aligned} S_v(n) &\doteq \boldsymbol{\Upsilon}_v^H(n)\boldsymbol{\Upsilon}_v(n) \\ &= \hat{\mathbf{x}}_v^H(n)\left(\mathbf{I}_4^H \otimes (\mathbf{C}^T\mathbf{H})^H\right)\left(\mathbf{I_4} \otimes \mathbf{C}^T\mathbf{H}\right)\hat{\mathbf{x}}_v(n) \\ &= \hat{\mathbf{x}}_v^H(n)\left(\mathbf{I}_4 \otimes \mathbf{W}\right)\hat{\mathbf{x}}_v(n). \end{aligned} \tag{23}$$

In (23), $\mathbf{I}_4$ and $\mathbf{W}$ are constant matrices $\forall n$. Hence the computation of the spectrum $S_v(n)$ for different windows of a DNA sequence needs only the evaluation of the Voss interleaved array $\hat{\mathbf{x}}_v(n)$.

3.2 Computing the DNA spectrum under general symbolic to numerical maps

Similar to the Voss representation case, any map $\mathbb{F} \longmapsto \mathbb{D}$ of γ sequences can be processed using the analysis of section 2. It directly follows that the ST-DFT and spectrum of a single sequence are given by

$$\begin{aligned} X_l(n) &= \mathbf{C}^T\mathbf{H}\ \hat{\mathbf{x}}_l(n), \\ S_l(n) &= \hat{\mathbf{x}}_l^H(n)\ \mathbf{W}\ \hat{\mathbf{x}}_l(n), \end{aligned}$$

where $l = 1, 2, \ldots, \gamma$. The array of γ $\mathbb{D}$-mapped ST-DFTs $\boldsymbol{\Upsilon}_d(n)$ is therefore given by

$$\begin{aligned} \boldsymbol{\Upsilon}_d(n) &= \left[\ X_1(n) \quad X_2(n) \quad \ldots \quad X_\gamma(n)\ \right]^T \\ &= \left(\mathbf{I}_\gamma \otimes \mathbf{C}^T\mathbf{H}\right)\hat{\mathbf{x}}_d(n). \end{aligned} \tag{24}$$

The $\mathbb{D}$-based DNA spectrum can easily be shown to be

$$S_d(n) = \hat{\mathbf{x}}_d^H(n)\left(\mathbf{I}_\gamma \otimes \mathbf{W}\right)\hat{\mathbf{x}}_d(n), \tag{25}$$

where the vector $\hat{\mathbf{x}}_d(n)$ of length γM is an array of the γ $\mathbb{D}$-mapped and interleaved windows starting at index n: $\hat{\mathbf{x}}_l(n)$, $\forall\ l = 1, 2, \ldots, \gamma$. It is clear that for every different map $\mathbb{F} \longmapsto \mathbb{D}$, a new interleaved windows array $\hat{\mathbf{x}}_d(n)$ has to be evaluated in order to compute a spectrum point $S_d(n)$. In this following, we introduce a different new approach to recompute (25) without updating $\hat{\mathbf{x}}_d(n)$ for every map. Basically, we derive a new expression for $S_d(n)$ in terms of $\hat{\mathbf{x}}_v(n)$ and a new constant matrix so that we incorporate the map dependance in the matrix part rather than the interleaved array part. In other words, since the map $\mathbb{F} \longmapsto \mathbb{V}$ is already well-defined, we use the map $\mathbb{V} \longmapsto \mathbb{D}$ to complete the chain $\mathbb{F} \longmapsto \mathbb{V} \longmapsto \mathbb{D}$ and hence find the spectrum $S_d(n)$. Consider the following affine transformation from Voss sequences to a general array of

$\mathbb{D}$-mapped sequences

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_\gamma(n) \end{bmatrix}_{\gamma\times 1} = \mathbf{A}_{\gamma\times 4} \begin{bmatrix} x_A(n) \\ x_C(n) \\ x_G(n) \\ x_T(n) \end{bmatrix}_{4\times 1} + \mathbf{b}_{\gamma\times 1},$$

where $\mathbf{A}_{\gamma\times 4}$ and $\mathbf{b}_{\gamma\times 1} = [b_1 \;\; b_2 \;\; \ldots \;\; b_\gamma]^T$ are constant possibly complex valued arrays. It follows that the array of the $\mathbb{D}$-mapped interleaved windows $\hat{\mathbf{x}}_d(n)$ can be written in terms of the array the Voss-mapped interleaved windows $\hat{\mathbf{x}}_v(n)$ in the following form

$$\hat{\mathbf{x}}_d(n)_{\gamma M\times 1} = (\mathbf{A}_{\gamma\times 4} \otimes \mathbf{I}_M)\, \hat{\mathbf{x}}_v(n)_{4M\times 1} + \hat{\mathbf{b}}_{\gamma M\times 1}, \tag{26}$$

where $\hat{\mathbf{b}}$ defined as

$$\hat{\mathbf{b}} = \left[\underbrace{b_1 \;..\; b_1}_{M} \;\; b_2 \;..\; b_2 \;\; \ldots \;\; \underbrace{b_\gamma \;..\; b_\gamma}_{M} \right]$$

is an array of γ M-element blocks, each block is M repetitions of one element of $\mathbf{b}$. Substituting for $\hat{\mathbf{x}}_d(n)$ in (24) results in a new formula for the array of $\mathbb{D}$-mapped ST-DFTs $\boldsymbol{\Upsilon}_d(n)$ into

$$\boldsymbol{\Upsilon}_d(n) = \left(\mathbf{I}_\gamma \otimes \mathbf{C}^T\mathbf{H}\right) \left[(\mathbf{A} \otimes \mathbf{I}_M)\, \hat{\mathbf{x}}_v(n) + \hat{\mathbf{b}} \right]. \tag{27}$$

An important result at this point is that the second term in $\boldsymbol{\Upsilon}_d(n)$ is actually equal to 0. This can be verified by reducing it into the following form

$$\left(\mathbf{I}_\gamma \otimes \mathbf{C}^T\mathbf{H}\right) \hat{\mathbf{b}} = \begin{bmatrix} \mathbf{C}^T\mathbf{H} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^T\mathbf{H} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{C}^T\mathbf{H} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \\ \vdots \\ \hat{\mathbf{b}}_\gamma \end{bmatrix}.$$

Recall that the sum of elements in $\mathbf{C}^T\mathbf{H}$ is equal to 0. Therefore, since $\hat{\mathbf{b}}_l$ is a constant vector, the product $\left(\mathbf{C}^T\mathbf{H}\right).\hat{\mathbf{b}}_l$ is equal to 0, $\forall\; l = 1, 2, \ldots, \gamma$ and hence

$$\left(\mathbf{I}_\gamma \otimes \mathbf{C}^T\mathbf{H}\right) \hat{\mathbf{b}} = \sum_{l=1}^{\gamma} \left(\mathbf{C}^T\mathbf{H}\right).\hat{\mathbf{b}}_l = 0. \tag{28}$$

The ST-DFTs array $\boldsymbol{\Upsilon}_d(n)$ can therefore be simplified using the Kronecker product identity (22) into

$$\begin{aligned} \boldsymbol{\Upsilon}_d(n) &= \left(\mathbf{I}_\gamma \otimes \mathbf{C}^T\mathbf{H}\right) (\mathbf{A} \otimes \mathbf{I}_M)\, \hat{\mathbf{x}}_v(n) \\ &= \left(\mathbf{A} \otimes \mathbf{C}^T\mathbf{H}\right) \hat{\mathbf{x}}_v(n). \end{aligned} \tag{29}$$

It follows that the $\mathbb{D}$-based DNA spectrum $S_d(n)$ is

$$\begin{aligned} S_d(n) &= \boldsymbol{\Upsilon}_d^H(n)\boldsymbol{\Upsilon}_d(n) \\ &= \hat{\mathbf{x}}_v^H(n)\left(\mathbf{A}\otimes\mathbf{C}^T\mathbf{H}\right)^H\left(\mathbf{A}\otimes\mathbf{C}^T\mathbf{H}\right)\hat{\mathbf{x}}_v(n) \\ &= \hat{\mathbf{x}}_v^H(n)\left(\mathbf{B}\otimes\mathbf{W}\right)\hat{\mathbf{x}}_v(n), \end{aligned} \tag{30}$$

where $\mathbf{B} \doteq \mathbf{A}^H\mathbf{A}$. Equation (30) indicates that when a certain symbolic to numeric mapping $\mathbb{F} \longmapsto \mathbb{D}$ is used, the DNA power spectrum $S_d(n)$ is completely defined in terms of the Voss-based interleaved array $\hat{\mathbf{x}}_v(n)$ along with constant matrices $\mathbf{W}$ and $\mathbf{B}$ which is a function of the transformation matrix $\mathbf{A}$ ($\mathbb{V} \longmapsto \mathbb{D}$). Note that if $\mathbf{A} = \mathbf{I}_4$ then $\mathbf{B} = \mathbf{I}_4$ at which (30) reduces to (23) which is the Voss-based spectrum case.

4 Invariance of the DNA spectrum under popular mappings

The results found in section 3 can be applied to some mappings that are widely used in the literature. In specific, by defining the corresponding transformation matrices $\mathbf{A}$ and $\mathbf{B}$ ($\mathbb{V} \longmapsto \mathbb{D}$), closed form expressions for $S_d(n)$ are obtained. Furthermore, for a number of mappings, we show that the $\mathbb{D}$-mapped spectrum $S_d(n)$ is in fact a scaled version of the Voss-based spectrum $S_v(n)$.

4.1 Four-to-four ($\gamma = 4$) representations

In this scheme, each Voss sequence is scaled by a possibly complex coefficient according to the following transformations matrices

$$\mathbf{A} = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & g & 0 \\ 0 & 0 & 0 & t \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} |a|^2 & 0 & 0 & 0 \\ 0 & |c|^2 & 0 & 0 \\ 0 & 0 & |g|^2 & 0 \\ 0 & 0 & 0 & |t|^2 \end{bmatrix},$$

where a, a, g, and t are real or complex coefficients used to scale $x_A(n)$, $x_C(n)$, $x_G(n)$, and $x_T(n)$ respectively. The corresponding array of ST-DFT's $\boldsymbol{\Upsilon}_d(n)$ is subsequently given by

$$\boldsymbol{\Upsilon}_d(n) = \left(\begin{bmatrix} a & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & g & 0 \\ 0 & 0 & 0 & t \end{bmatrix} \otimes \mathbf{C}^T\mathbf{H}\right)\hat{\mathbf{x}}_v(n),$$

and the DNA spectrum $S_d(n)$ is

$$S_d(n) = \hat{\mathbf{x}}_v^H(n)\left(\begin{bmatrix} |a|^2 & 0 & 0 & 0 \\ 0 & |c|^2 & 0 & 0 \\ 0 & 0 & |g|^2 & 0 \\ 0 & 0 & 0 & |t|^2 \end{bmatrix} \otimes \mathbf{W}\right)\hat{\mathbf{x}}_v(n).$$

Now, we extend this result to certain transformations where numeric values of the scale factors a, a, g, and t are specified.

§ *Tetrahedral mapping.* The so-called tetrahedral representation has been proposed in [13, 29]. In this mapping scheme, the four nucleotides are represented by four equal length vectors oriented towards the corners of a tetrahedron. Projecting the basic tetrahedron on a plane will reduce the dimensionality of the representation to two. This mapping can be defined by the mapping matrix

$$\mathbf{A} = \begin{bmatrix} 1+j & 0 & 0 & 0 \\ 0 & -1+j & 0 & 0 \\ 0 & 0 & -1-j & 0 \\ 0 & 0 & 0 & 1-j \end{bmatrix}.$$

It can be easily seen that in this case: $|a| = |c| = |g| = |t| = \sqrt{2}$ which implies that $\mathbf{B} = 2\mathbf{I_4}$. The corresponding DNA spectrum is

$$S_d(n) = 2\hat{\mathbf{x}}_v^H(n)\,(\mathbf{I_4} \otimes \mathbf{W})\,\hat{\mathbf{x}}_v(n) = 2S_v(n). \tag{31}$$

Since $\mathbf{B} = \alpha\mathbf{I}_4$ ($\alpha = 2$), the tetrahedral-based DNA spectrum is a scaled version of the Voss-based spectrum.

§ *Quaternion mapping.* A more involved step is to replace the complex number set of the tetrahedral mapping with its algebraic generalization, the set of quaternions. Quaternions have been used to map DNA sequences $\mathbb{F} \longmapsto \mathbb{H}$ [30] and are simply defined as hypercomplex numbers given by $p \in \mathbb{H} = \{a + bi + cj + dk | a, b, c, d \in \mathbb{R}\}$, where i, j, k are complex coefficients such that $i^2 = j^2 = k^2 = ijk = -1$ and $|p| = \sqrt{pp^*} = \sqrt{a^2 + b^2 + c^2 + d^2}$. The transformation matrix is given by

$$\mathbf{A} = \begin{bmatrix} i+j+k & 0 & 0 & 0 \\ 0 & i-j-k & 0 & 0 \\ 0 & 0 & -i-j+k & 0 \\ 0 & 0 & 0 & -i+j-k \end{bmatrix}.$$

In this case, $|a| = |c| = |g| = |t| = \sqrt{3}$, $\mathbf{B} = 3\mathbf{I_4}$. The corresponding DNA spectrum is

$$S_d(n) = 3\hat{\mathbf{x}}_v^H(n)\,(\mathbf{I_4} \otimes \mathbf{W})\,\hat{\mathbf{x}}_v(n) = 3S_v(n) \tag{32}$$

§ *Higher order mappings.* An alternative Quaternion transformation is given by $\mathbf{A} = diag(1 + i + j + k, 1 + i - j - k, 1 - i - j + k, 1 - i + j - k)$, which results in $\mathbf{B} = 4\mathbf{I_4}$ and consequently $S_d(n) = 4S_v(n)$. In general, for a complex representation system with η dimensions and equal amplitude coefficients: $\mathbf{B} = \eta\mathbf{I_4}$ and hence the spectrum $S_d(n) = \eta S_v(n)$.

4.2 Four-to-three ($\gamma = 3$) mappings

In order to reduce the DNA spectrum computational cost, several mappings have been proposed with smaller numbers of sequences.

§ *$\mathcal{Z}$-curve mapping.* One such important symbolic-to-numeric map is the $\mathcal{Z}$-curve mapping [24], which is a unique 3-dimensional curve representation whose sequences have values 1 and -1. One advantage of the $\mathcal{Z}$-curve mapping is that each of its three sequences has a biological interpretation. This scheme is given by

$$\begin{bmatrix} x(n) \\ y(n) \\ z(n) \end{bmatrix} = 2 \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_A(n) \\ x_C(n) \\ x_G(n) \\ x_T(n) \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Therefore, the transformation matrices are

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 12 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix}.$$

Matrix $\mathbf{B}$ in this case can be written as

$$\begin{aligned} \mathbf{B} &= 4 \left(\mathbf{I}_4 + \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \right) \\ &= 4 \left(\mathbf{I}_4 + \mathbf{B}_1 + \mathbf{B}_2 \right). \end{aligned}$$

Note that the term involving $\mathbf{B}_1$ in $S_d(n)$ can be manipulated into

$$\begin{aligned} S_d(n)|_{\mathbf{B}_1} &= \hat{\mathbf{x}}_v^H(n) \left(\mathbf{B}_1 \otimes \mathbf{W} \right) \hat{\mathbf{x}}_v(n) \\ &= 4\hat{\mathbf{x}}_v^H(n) \begin{bmatrix} \mathbf{W} & \mathbf{W} & \mathbf{W} & \mathbf{W} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_A(n) \\ \hat{\mathbf{x}}_C(n) \\ \hat{\mathbf{x}}_G(n) \\ \hat{\mathbf{x}}_T(n) \end{bmatrix} = 4\hat{\mathbf{x}}_v^H(n) \begin{bmatrix} \mathbf{W} \left(\sum_{l \in \mathbb{F}} \hat{\mathbf{x}}_l(n) \right) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \end{aligned}$$

Recall from (17) that $\sum_{l \in \mathbb{F}} \hat{\mathbf{x}}_l(n) = [1 \ 1 \ldots 1]^T$. Take also into consideration that the sum of elements of any row or column in $\mathbf{W}$ is equal to 0. This implies that $\mathbf{W} \left(\sum_{l \in \mathbb{F}} \hat{\mathbf{x}}_l(n) \right) = \mathbf{0}$, at which it is easy to see that $S_d(n)|_{\mathbf{B}_1} = 0$. Similarly, $S_d(n)|_{\mathbf{B}_2} = 0$. Therefore, only the first term in $\mathbf{B}$ contributed to $S_d(n)$ at which the $\mathcal{Z}$-curve mapped DNA spectrum is a scaled version of the Voss-based DNA spectrum

$$S_d(n) = \hat{\mathbf{x}}_v^H(n) \left(4\mathbf{I_4} \otimes \mathbf{W} \right) \hat{\mathbf{x}}_v(n) = 4S_v(n). \tag{33}$$

This ratio is consistent with the result we first derived in [24] for $R = 3$, but is now shown to be general for any value of R. We are now ready to state an important result.

Theorem. Necessary and Sufficient condition for the invariance of the DNA spectrum. Consider the following affine transformation from Voss sequences to a general array of $\mathbb{D}$-mapped sequences

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_\gamma(n) \end{bmatrix}_{\gamma\times 1} = \mathbf{A}_{\gamma\times 4} \begin{bmatrix} x_A(n) \\ x_C(n) \\ x_G(n) \\ x_T(n) \end{bmatrix}_{4\times 1} + \mathbf{b}_{\gamma\times 1},$$

where $\mathbf{A}_{\gamma\times 4}$ and $\mathbf{b}_{\gamma\times 1} = [b_1 \;\; b_2 \;\; \ldots \;\; b_\gamma]^T$ are constant possibly complex valued arrays. Define the 4×4 matrix $\mathbf{B} = \mathbf{A}^{\mathbf{H}}\mathbf{A}$. The DNA spectrum is invariant under this map, i.e., $S_d(n) = \alpha S_v(n)$ if the transformation matrix $\mathbf{B}$ can be written as $\mathbf{B} = \alpha\mathbf{I}_4 + \sum_i \mathbf{B}_i$ where $\mathbf{B}_i$ holds constant rows and/or constant columns $\forall\; i$. The proof follows by simply observing that if $\mathbf{B}_i$ has constant rows and/or constant columns, then $S_d(n)|_{\mathbf{B}_i} = 0$. We remind the reader at this point that the vector $\mathbf{b}_{\gamma\times 1}$ has no bearing on the invariance of the DNA spectrum.

§ *Simplex mapping.* The simplex mapping is essentially another tetrahedron structured mapping that aims to eliminate the computational redundancy. Its transformations matrices are

$$\mathbf{A} = \frac{1}{3}\begin{bmatrix} 0 & -\sqrt{2} & -\sqrt{2} & 2\sqrt{2} \\ 0 & \sqrt{6} & -\sqrt{6} & 0 \\ 3 & -1 & -1 & -1 \end{bmatrix},\; \mathbf{B} = \frac{1}{3}\begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix}.$$

Matrix $\mathbf{B}$ in this case can be written as

$$\mathbf{B} = \left(\frac{4}{3}\right)\left(\mathbf{I}_4 - \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}\right) = \left(\frac{4}{3}\right)(\mathbf{I}_4 + \mathbf{B}_1).$$

Similar to the $\mathcal{Z}$-curve case, $S_d(n)|_{\mathbf{B}_1} = 0$. It follows that the simplex-based DNA spectrum is also a scaled version of the Voss-based spectrum, and is given by

$$S_d(n) = \hat{\mathbf{x}}_v^H(n)\left(\frac{4}{3}\mathbf{I_4}\otimes\mathbf{W}\right)\hat{\mathbf{x}}_v(n) = \left(\frac{4}{3}\right)S_v(n). \tag{34}$$

This ratio is consistent with the result in [31] which was limited to direct DFT and is now shown to be extended to M/R ST-DFT with any value of R.

4.3 Four-to-two ($\gamma = 2$) mappings

Pairing couples of nucleotides together was proposed in the literature in order to exploit certain biological features in addition to complexity reduction. For example, it was suggested that exons are rich in nucleotides

C and G, while introns have more A and T [29]. This claim inspired the transformation

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & -1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

It is obvious that the DNA spectrum in this case can be simplified to

$$S_d(n) = \hat{\mathbf{x}}_v^H(n) \begin{bmatrix} \mathbf{W} & \mathbf{0} & \mathbf{0} & \mathbf{W} \\ \mathbf{0} & \mathbf{W} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} & \mathbf{W} & \mathbf{0} \\ \mathbf{W} & \mathbf{0} & \mathbf{0} & \mathbf{W} \end{bmatrix} \hat{\mathbf{x}}_v(n), \tag{35}$$

which obviously is not a scaled version of $S_v(n)$ since $\mathbf{B}$ in this case can not be written as $\alpha\mathbf{I}_4 + \sum_i \mathbf{B}_i$ where $\mathbf{B}_i$ holds constant rows and/or constant columns $\forall\ i$.

4.4 Four-to-one ($\gamma = 1$) mappings

Single sequence representations can be generated by assigning each nucleotide a certain coefficient [4, 13] in order to keep the single sequence structure using the transformation array and matrix

$$\mathbf{A} = \begin{bmatrix} a & c & g & t \end{bmatrix}, \mathbf{B} = \begin{bmatrix} |a|^2 & a^*c & a^*g & a^*t \\ c^*a & |c|^2 & c^*g & c^*t \\ g^*a & g^*c & |g|^2 & g^*t \\ t^*a & t^*c & t^*g & |t|^2 \end{bmatrix}.$$

Note that the coefficients chosen for the tetrahedral, quaternion, and paired coupled mappings can be reused along with the single sequence formulation. For example, the paired couples case can be reformulated in a single sequence of 1's and -1's using $\mathbf{A} = \begin{bmatrix} -1 & 1 & 1 & -1 \end{bmatrix}$ and

$$\mathbf{B} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

at which the DNA spectrum is

$$S_d(n) = \hat{\mathbf{x}}_v^H(n) \begin{bmatrix} \mathbf{W} & -\mathbf{W} & -\mathbf{W} & \mathbf{W} \\ -\mathbf{W} & \mathbf{W} & \mathbf{W} & -\mathbf{W} \\ -\mathbf{W} & \mathbf{W} & \mathbf{W} & -\mathbf{W} \\ \mathbf{W} & -\mathbf{W} & -\mathbf{W} & \mathbf{W} \end{bmatrix} \hat{\mathbf{x}}_v(n).$$

Similar to the previous case, $S_d(n)$ is not a scaled version of $S_v(n)$.

Experimental Verification. To briefly verify the results of this section experimentally, we apply Equation (30) to real DNA sequences, when the Voss, tetrahedral, quaternion, Z-curve, and simplex maps are

employed. For comparison with previous work, we consider first the DNA sequence F56F11.4 in the *C. elegans* chromosome III. This sequence is 8060 nucleotides and has been used as a benchmark by many researchers [13] to extract the periodicity component at $R = 3$. The DNA spectra at $R = 3$ are shown in Figure 8 for the five former mappings, and are obviously related by the constant scale factors derived earlier in the section which clearly verifies our results. Although we lack the space for more general simulations, it is important to state that all the spectra relations are maintained experimentally at other values of R associated with higher order periodicities.

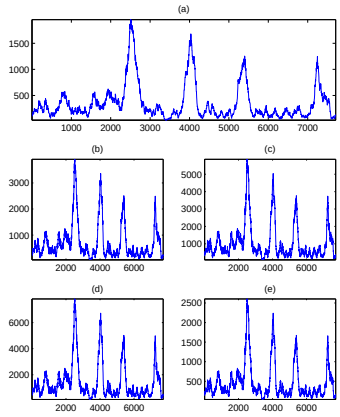


Figure 8: DNA spectrum $S_d(n)$ at $R = 3$ of the DNA sequence F56F11.4 when (a) Voss, (b) tetrahedral, (c) quaternion, (d) Z-curve, and (e) simplex mappings are used.

For generality purposes, we test two more sequences extracted from the well known Burset-Guigo database [32]. In specific, DNA spectra at $R = 3$ of the zeta globin gene (ECZGL2) of length 1563, and the Alouatta seniculus epsilon-globin gene (ALOEGLOBIM) of length 1691 are shown in Figure 9 and Figure 10, respectively, for the five former mappings. It can be seen that the relations are still preserved.

5 Alternative Measures of DNA periodicities

Alternative DNA periodicity measures using fast data transforms [33–35], wavelets [36], and finite impulse response (FIR) digital filters [25, 37] were recently proposed to improve the detection performance of these periodicities. However, each method was obtained separately from the other using seemingly a different approach. In this section, we show that our proposed framework can systematically generate all these results by simply changing a number of matrices. It therefore provides a *generic unified framework* for generating alternative measures of DNA periodicities. For example, we can re-express the matrices **D** and

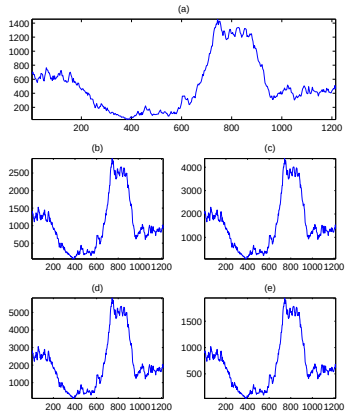


Figure 9: DNA spectrum $S_d(n)$ at $R = 3$ of the DNA sequence ECZGL2 when (a) Voss, (b) tetrahedral, (c) quaternion, (d) Z-curve, and (e) simplex mappings are used.

W in terms of general digital filters and use these filters to modify (30) in order to generate new spectrum formulas. Furthermore, using symmetry based decompositions of **D** and **W**, we simplify (30) into a formula with low computational complexity.

5.1 Modified Periodicity Measures

Recall from section 2 that matrix **W** is given by

$$\mathbf{W} \quad = \quad \mathbf{H}^H\mathbf{D}\mathbf{H} = (\mathbf{I}_R \otimes \mathbf{h}^T)^H\mathbf{C}^\star\mathbf{C}^T(\mathbf{I}_R \otimes \mathbf{h}^T).$$

Obviously, **W** is completely defined by the real array **h** and the generally complex array **C**. Note that **h** and **C** can be viewed as the impulse responses of two FIR filters defined by the z-transforms $H(z)$ and $C(z)$.

5.1.1 Updating the real filter **h**

The FIR filter $H(z)$ is the standard rectangular window filter and has a low pass frequency response with a -13 dB attenuation. To improve its filtering performance, we can use a more general FIR filter, denoted by $\tilde{H}(z)$ and expressed as

$$\tilde{H}(z) = h_0 + h_1 z^{-1} + \cdots + h_{\frac{M}{R}} z^{-\frac{M}{R}},$$

which is the $\mathcal{Z}$-transform of the general array $\tilde{\mathbf{h}}$ given by

$$\tilde{\mathbf{h}} = \left[h_0 \quad h_1 \quad \ldots \quad h_{\frac{M}{R}}\right].$$

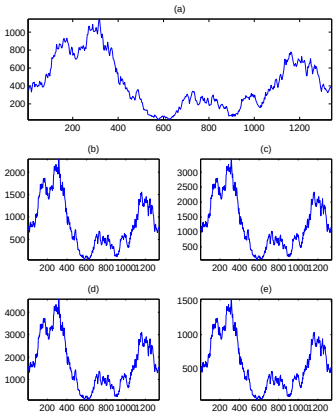
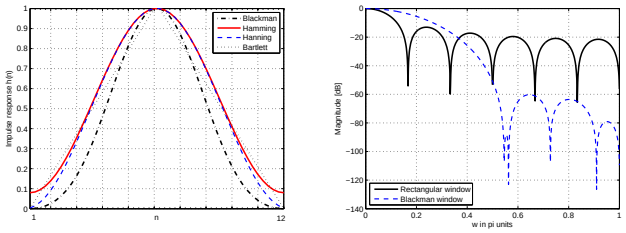


Figure 10: DNA spectrum $S_d(n)$ at $R = 3$ of the DNA sequence ALOEGLOBIM when (a) Voss, (b) tetrahedral, (c) quaternion, (d) Z-curve, and (e) simplex mappings are used.

From a signal processing perspective, achieving better performance can be obtained by replacing the rectangular window with another one, $\tilde{H}(z)$, that has slightly wider main lobes but much more attenuated side lobes, as shown in Table 3. The impulse responses of such windows are depicted in Figure 11(a) for $R = 8$ and $M = 96$. Better harmonics characterization can be achieved by giving each nucleotide position within the window a relative weight in contrast to the rectangular where equal weighting is given to all nucleotides. It turns out that the Blackman window has the best main-to-first side lobe attenuation behavior as shown in Figure 11(b) compared to the rectangular window case and therefore provides the best smoothing of the DNA spectrum.



(a) Impulse responses of standard FIR windows

(b) Magnitude response of rectangular and Blackman windows

Figure 11: Comparison between standard FIR windows showing (a) impulse response (b) magnitude response, when $R = 8$ and $M = 96$.

By replacing $\mathbf{h}$ with $\tilde{\mathbf{h}}$, the matrix $\mathbf{H}$ can be in turn expressed as

$$\tilde{\mathbf{H}} = \begin{bmatrix} h_0\ h_1\ ..\ h_{M/R-1} & \dots & 0\ 0\ ..\ 0 \\ \vdots & \ddots & \vdots \\ 0\ 0\ ..\ 0 & \dots & \underbrace{h_0\ h_1\ ..\ h_{M/R-1}}_{M/R} \end{bmatrix},$$

and the complex row vector $\mathbf{C}^T\tilde{\mathbf{H}}$ is now given by

$$\mathbf{C}^T\tilde{\mathbf{H}} = \begin{bmatrix} \underbrace{h_0\ ..\ h_{\frac{M}{R}-1}}_{M/R} & \dots & \underbrace{h_0 e^{-j2\pi\frac{R-1}{R}}\ ..\ h_{\frac{M}{R}-1}e^{-j2\pi\frac{R-1}{R}}}_{M/R} \end{bmatrix}.$$

It can be easily seen that the sum of elements in $\mathbf{C}^T\tilde{\mathbf{H}}$ is still equal to zero as was the case for $\mathbf{C}^T\mathbf{H}$. Consequently, it follows that the sum of any row or column in $\tilde{\mathbf{W}} = \tilde{\mathbf{H}}^H\mathbf{D}\tilde{\mathbf{H}}$ is still equal to zero. This is a fundamental result which, in turn, implies that all the derivations of section 3 are still the same even when $\tilde{\mathbf{h}}$ replaces $\mathbf{h}$. In particular, the $\mathbb{V}$-based DNA spectrum $\tilde{S}_v(n)$ and the $\mathbb{D}$-based one $\tilde{S}_d(n)$ can be stated as

$$\tilde{S}_v(n) = \hat{\mathbf{x}}_v^H(n)(\mathbf{I}_4 \otimes \tilde{\mathbf{W}})\hat{\mathbf{x}}_v(n),\ \ \tilde{S}_d(n) = \hat{\mathbf{x}}_v^H(n)(\mathbf{B} \otimes \tilde{\mathbf{W}})\hat{\mathbf{x}}_v(n). \tag{36}$$

Moreover, all the mathematical relations derived in section 3 between the $\mathbb{D}$-based spectrum and the Voss-based one are all still valid even when $\mathbf{h}$ is replaced by $\tilde{\mathbf{h}}$.

Experimental Verification. To experimentally verify this result, we consider finding the DNA spectrum $\tilde{S}_d(n)$ of the 3 DNA sequences used in the previous section when $\tilde{\mathbf{h}}$ is set to a Blackman window. The relations between the spectra when using the Voss, tetrahedral, quaternion, Z-curve, and simplex mappings are still the same as shown in Figure 12, Figure 13, and Figure 14.

5.1.2 Updating the complex filter $\mathbf{C}$

Similar to $H(z)$, the FIR filter $C(z)$ can be replaced by a more sophisticated filter $\tilde{C}(z)$ expressed as

$$\tilde{C}(z) = C_0 + C_1 z^{-1} + \cdots + C_{R-1} z^{-(R-1)},$$

which is the $\mathcal{Z}$-transform of the general array $\tilde{\mathbf{C}}$ given by

$$\tilde{\mathbf{C}} = [C_0\ \ C_1\ \ \dots\ \ C_{R-1}].$$

Note that, in this case, the elements in array $\tilde{\mathbf{C}}$ do not necessarily add to zero anymore. Consequently, the sum of elements in any row or any column in $\tilde{\mathbf{D}} = \tilde{\mathbf{C}}^\star\tilde{\mathbf{C}}^T$ or $\tilde{\mathbf{W}} = \mathbf{H}^H\tilde{\mathbf{D}}\mathbf{H}$ is not necessarily zero. We

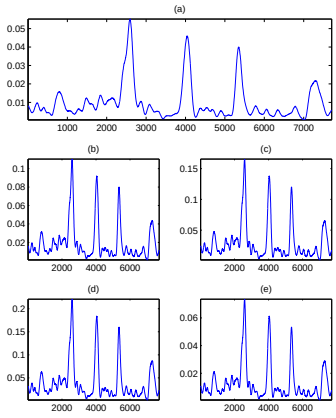


Figure 12: DNA spectrum $\tilde{S}_d(n)$ with $\tilde{\mathbf{h}}$ set to a Blackman window at $R = 3$ of the DNA sequence F56F11.4 when (a) Voss, (b) tetrahedral, (c) quaternion, (d) Z-curve, and (e) simplex mappings are used.

also note that unlike the case of $\tilde{\mathbf{h}}$, using $\tilde{\mathbf{C}}$ instead of $\mathbf{C}$ keeps the spectrum formulas in (36) correct but does not preserve the mathematical relations between the different $\mathbb{D}$-mapped spectra and the Voss-based spectrum.

5.1.3 Joint Optimization of $\tilde{\mathbf{h}}$ *and* $\tilde{\mathbf{C}}$

It should be clear at this point that better DNA harmonics detection performance can be potentially achieved through a joint "optimization" of $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{C}}$. For example, a learning paradigm can be used with a least-mean-square (LMS) criterion to find the optimal set, $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{C}}$. Alternatively, a biologically induced criterion can yield a substantial boost in performance but it is not clear which criterion to use. This interesting but challenging research topic is however outside the scope of this paper and will not be further pursued here.

Example. Standard discrete time transforms have been proposed to replace the ST-DFT in the periodicity detection problem. In particular, the short time discrete cosine transform (ST-DCT), sine transform (ST-DST), and Hartley transform (ST-DHT) were introduced and analyzed for this purpose [33]. In this example, we show that these three transforms fit naturally within our proposed analysis when the two arrays $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{C}}$ are adjusted correctly for each case. Although these standard transforms are not optimized for certain data sets, they can serve as preliminary tests for better periodicity detection. In [33], the short time DFT,

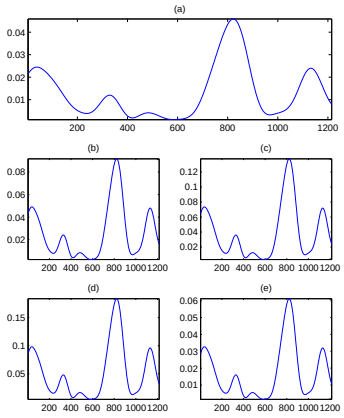


Figure 13: DNA spectrum $\tilde{S}_d(n)$ with $\tilde{\mathbf{h}}$ set to a Blackman window at $R = 3$ of the DNA sequence ECZGL2 when (a) Voss, (b) tetrahedral, (c) quaternion, (d) Z-curve, and (e) simplex mappings are used.

DCT, DST, and DHT at $k = M/R$ where shown to be given by

$$X^{(t)}(n) = \sum_{r=0}^{R-1} C_r^{(t)} \sum_{m=r,r+R,\ldots}^{\frac{M}{R}-1} x(n + mR + r)h^{(t)}(m), \tag{37}$$

where $t \in \{f, c, s, h\}$ indicates Fourier, cosine, sine, and Hartley transforms, respectively, $C_r^{(t)} = a^{(t)}e^{j\theta_r^{(t)}} + b^{(t)}e^{-j\theta_r^{(t)}}$ are possibly complex coefficients, and $h^{(t)}(m) = (\alpha^t)^m$. Values of the parameters α, a, b, and θ_r for every transform are adjusted according to Table 4. For illustration, setting $\alpha = 1$, $a = 1$, $b = 0$, and $\theta_r = -2\pi r/R$ in (37) results in the ST-DFT case. An efficient implementation to calculate Equation (37) is shown in Figure 15 which generalizes Figure 3. This model provides a general framework that encapsulates the computation of the short-time Fourier, cosine, sine, and Hartley transforms at frequency point $k = M/R$. Therefore, the same matrix-based analysis of sections 2 and 3 can be used. Matrix $\mathbf{W}$ will be updated into

$$\begin{aligned} \tilde{\mathbf{W}} &= \tilde{\mathbf{H}}^H \tilde{\mathbf{D}} \tilde{\mathbf{H}} \\ &= (\mathbf{I}_R \otimes \tilde{\mathbf{h}}^T)^H \tilde{\mathbf{C}}^\star \tilde{\mathbf{C}}^T (\mathbf{I}_R \otimes \tilde{\mathbf{h}}^T), \end{aligned}$$

and therefore the $\mathbb{D}$-based DNA spectrum $\tilde{S}_d(n)$ when one of the ST- DFT, DCT, DST, or DHT is employed can be stated as

$$\tilde{S}_d(n) = \hat{\mathbf{x}}_v^H(n)(\mathbf{B} \otimes \tilde{\mathbf{W}})\hat{\mathbf{x}}_v(n), \tag{38}$$

where the values of $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{C}}$ are adjusted according to Table 5. Note that similar to the Fourier case, the sum of elements in $\tilde{\mathbf{C}}$ for the cosine and Hartley transforms cases is equal to zero. Therefore, under these

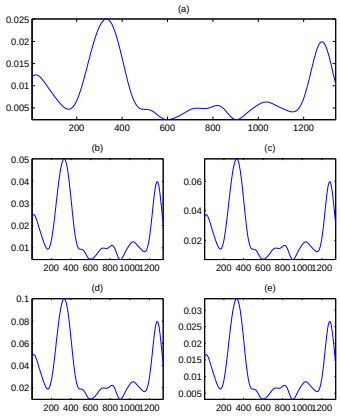


Figure 14: DNA spectrum $\tilde{S}_d(n)$ with $\tilde{\mathbf{h}}$ set to a Blackman window at $R = 3$ of the DNA sequence ALOE-GLOBIM when (a) Voss, (b) tetrahedral, (c) quaternion, (d) Z-curve, and (e) simplex mappings are used.

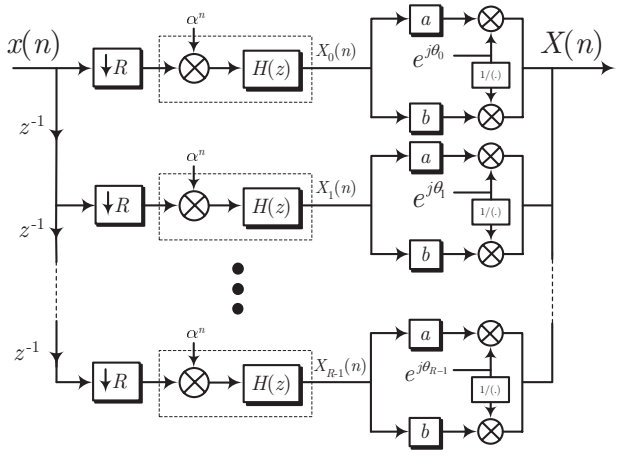


Figure 15: A general multirate DSP structure to compute the short-time DFT, DCT, DST and DHT.

two cases, the relations between different $\mathbb{D}$-based DNA spectra and the $\mathbb{V}$-based DNA spectrum are still the same as given in section 3. ■

At this point, it can be concluded that the $\mathbb{D}$-based DNA spectrum $\tilde{S}_d(n)$ is completely defined in terms of the Voss-based array of interleaved windows $\hat{\mathbf{x}}_v(n)$, the $\mathbb{V} \longmapsto \mathbb{D}$ mapping matrix $\mathbf{A}$, the real array $\tilde{\mathbf{h}}$, and the generally complex array $\tilde{\mathbf{C}}$. This conclusion is summarized in Figure 16.

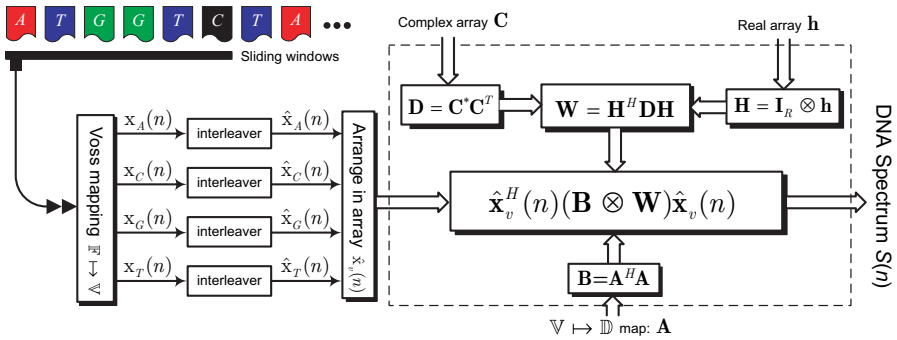


Figure 16: A DSP structure to compute the modified $\mathbb{D}$-based DNA spectrum $\tilde{S}_d(n)$. The Voss-based array of interleaved windows $\hat{\mathbf{x}}_v(n)$, the $\mathbb{V} \longmapsto \mathbb{D}$ mapping matrix $\mathbf{A}$, the real array $\mathbf{h}$, and the generally complex array $\tilde{\mathbf{C}}$ are the system design parameters.

5.2 A Real Approach for the Spectrum Computation

A real computationally-efficient alternative for the evaluation of $S_d(n)$ can be found by observing the special properties of the circulant/toeplitz matrix $\mathbf{D}$ or equivalently the block matrix $\mathbf{W}$. We use the fact that for a generally-complex matrix $\mathbf{Q}$: $y^H \mathbf{Q} y = 0$, $\forall y \in \mathbb{R}$, if $\mathbf{Q}$ is an antisymmetric matrix. We start by splitting $\mathbf{D}$ into its symmetric and antisymmetric parts

$$\mathbf{D} = \underbrace{\frac{1}{2}\left(\mathbf{D} + \mathbf{D}^T\right)}_{symmetric} + \underbrace{\frac{1}{2}\left(\mathbf{D} - \mathbf{D}^T\right)}_{antisymmetric} = \mathbf{D}_s + \mathbf{D}_{as},$$

where $\mathbf{D}_s$ is a circulant and Toeplitz real $R \times R$ matrix given by

$$\mathbf{D}_s = \begin{bmatrix} 1 & 2\cos\frac{2\pi}{R} & \dots & 2\cos\frac{2\pi(R-1)}{R} \\ 2\cos\frac{2\pi}{R} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 2\cos\frac{2\pi}{R} \\ 2\cos\frac{2\pi(R-1)}{R} & \dots & 2\cos\frac{2\pi}{R} & 1 \end{bmatrix},$$

and $\mathbf{D}_{as}$ is a circulant and Toeplitz complex $R \times R$ matrix given by

$$\mathbf{D}_{as} = 2j \begin{bmatrix} 0 & -\sin\frac{2\pi}{R} & \dots & -\sin\frac{2\pi(R-1)}{R} \\ \sin\frac{2\pi}{R} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\sin\frac{2\pi}{R} \\ \sin\frac{2\pi(R-1)}{R} & \dots & \sin\frac{2\pi}{R} & 0 \end{bmatrix}.$$

Substituting for $\mathbf{D}$ in (15), we get a simple form of the spectrum $S(n)$

$$\begin{aligned} S(n) &= \hat{\mathbf{x}}^H(n)\, \mathbf{H}^H \mathbf{D} \mathbf{H}\, \hat{\mathbf{x}}(n) \\ &= \hat{\mathbf{x}}^H(n)\, \mathbf{H}^H \left(\mathbf{D}_s + \mathbf{D}_{as}\right) \mathbf{H}\, \hat{\mathbf{x}}(n) \\ &= \hat{\mathbf{x}}^H(n)\, \mathbf{W}_s\, \hat{\mathbf{x}}(n), \end{aligned} \tag{39}$$

where $y^H \mathbf{D}_{as} y = 0, \forall l \in \mathbb{F}$, $y = \mathbf{H}\ \hat{\mathbf{x}}(n)$. The block matrix

$$\mathbf{W}_s \doteq \mathbf{H}^H \mathbf{D}_s \mathbf{H} = \frac{1}{2}\mathbf{H}^H \left(\mathbf{D} + \mathbf{D}^T\right) \mathbf{H}$$

is an $R \times R$ matrix of $\frac{M}{R} \times \frac{M}{R}$ blocks. Using (39) to update the DNA spectrum (19), $S_v(n)$ simplifies into

$$S_v(n) \quad = \quad \sum_{l \in \mathbb{F}} \hat{\mathbf{x}}_l^H(n)\ \mathbf{W}_s\ \hat{\mathbf{x}}_l(n). \tag{40}$$

Following the same analysis of section 3, (40) can be easily manipulated into a more elegant completely real form given by

$$S_v(n) = \hat{\mathbf{x}}_v^H(n)\,(\mathbf{I}_4 \otimes \mathbf{W}_s)\,\hat{\mathbf{x}}_v(n),$$

or more generally, (30) can be updated into

$$S_d(n) = \hat{\mathbf{x}}_v^H(n)\,(\mathbf{B} \otimes \mathbf{W}_s)\,\hat{\mathbf{x}}_v(n), \tag{41}$$

which provides a completely real approach for the computation of the $\mathbb{D}$-mapped spectrum $S_d(n)$. Note that all results and different spectra relations in section 3 still hold when $\mathbf{W}_s$ replaces $\mathbf{W}$ as in (41).

Computational complexity comparison. To quantify the computational credit of this real approach, we compare the complexity of (39) to that of (15) of a single discrete time sequence. Since $\hat{\mathbf{x}}(n)$ can be complex as well according to the mapping used, we find the number of real multiplications and additions needed to evaluate (39) when each of $\hat{\mathbf{x}}(n)$ and $\mathbf{W}$ is either real or complex, as given in Table 6. Recall that the multiplication of the complex numbers x and y, where $x = a + jb$ and $y = c + jd$ requires the computation of $ac - bd$ and $ad + bc$, which requires four real multiplications and two real additions.

Example. For illustration, we evaluate the spectrum $S_v(n)$ using $\mathbf{W}_s$ when $R = 3$, and compare the result to the formula derived in [38]. In specific, we use (40) to find the spectrum $S(n)$ as follows

$$\begin{aligned} S_v(n) \quad &= \quad \sum_{l \in \mathbb{F}} \hat{\mathbf{x}}_l^H(n) \mathbf{H}^H \mathbf{D}_s \mathbf{H} \hat{\mathbf{x}}_l(n) = \sum_{l \in \mathbb{F}} \mathbf{\Gamma}_l^H(n) \mathbf{D}_s\ \mathbf{\Gamma}_l(n) \\ &= \quad \sum_{l \in \mathbb{F}} [X_{l0} \quad X_{l1} \quad X_{l2}] \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} X_{l0} \\ X_{l1} \\ X_{l2} \end{bmatrix}. \end{aligned}$$

Expanding and completing the square, it follows that

$$\begin{aligned} S_v(n) &= \sum_{l\in\mathbb{F}}[X_{l0}^2(n) + X_{l1}(n)(X_{l1}(n) - X_{l0}(n)) \\ &+ X_{l2}(n)(X_{l2}(n) - X_{l0}(n) - X_{l1}(n))] \\ &= \frac{1}{2}\sum_{l\in\mathbb{F}}\sum_{r=0}^{2}(X_{lr}(n) - X_{lq}(n))^2, \end{aligned} \tag{42}$$

where $q = (r+1) \bmod 3$. The matrix-based DNA spectrum formula in (42) is consistent with the result derived using a different approach in [38]. ■

6 Concluding Remarks

In this paper, we have introduced a matrix based framework for locating hidden DNA periodicities using spectral analysis techniques that are invariant to the choice of the symbolic to numerical map. The primary advantage of the presented approach over some of the previous work is the decomposition of the spectrum expression into key matrices whose values can be set *independently from each other*. Each matrix represents one of the essential components involved in the computation of the spectrum such as the symbolic to numerical map, the data transform, and the filtering scheme. The above framework is derived under the assumption that the symbolic to numerical map can be obtained from the Voss representation using an affine transformation. This assumption is however quite loose given that most (if not all) of the proposed maps in the literature satisfy this requisite. Using the new framework, we have then shown that the DNA spectrum expression is invariant under these maps. We have also derived a necessary and sufficient condition for the invariance of the DNA spectrum in terms of the affine transformation matrix $\mathbf{A}$ (the $\mathbf{b}$ vector in the affine transformation does not affect the DNA spectrum). This condition can serve as the basis for generating novel symbolic to numerical map that preserve the DNA spectrum expression. Finally, in the latter sections of the paper, we have shown the potential of using different filtering schemes e.g. windows other than the rectangular one as well as alternate fast data transforms e.g. the DCT, DST, and the Hartley transform. A number of simulation results that verify the findings of this paper and a brief quantitative analysis of the computational complexity of the new approach were given in the same sections. Future research work would consider the optimization of the different building blocks, namely the symbolic to numerical map, the data transform, and the filtering scheme. This, in turn, requires a deep understanding of the biological significance of different DNA periodicities in order to set up a meaningful objective function and appropriate constraints. Ultimately, the framework proposed here can be incorporated in a more sophisticated system to study the complex structure of genomic sequences and understand the functionality of its various components. Finally,

this efficient framework can be extended to the analysis of other types of symbolic sequences of various limited alphabets, either biological sequences (such as protein sequences) or even non-biological ones.

References

1. Benson G: **Tandem Repeat Finder: A program to Analyze DNA Sequences**. *Nucleic Acids Research* 1999, **27**(2):573–580.
2. Butler J: *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press 2003.
3. Cummings CA, Relman DA: **Microbial forensics: cross-examining pathogens**. *Science* 2002, **296**:1976–1979.
4. Ramachandran P, Lu W, Antoniou A: **Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA**. *IEEE Transactions on Biomedical Engineering* 2012, **59**(6):1598–1609.
5. Strachan T, Read AP: *Human Molecular Genetics*. John Wiley and Sons 1999.
6. Smit AF: **The origin of interspersed repeats in the human genome**. *Curr. Opin. Genet. Dev.* 1996, **6**:743–748.
7. Rubinsztein DC, Hayden MR: *Analysis of triplet repeat disorders*. Bios Scientific Pub. 1999.
8. Gupta R, Mittal A, Gupta S: **An efficient Algorithm to Detect Palindromes in DNA Sequences using Periodicity Transform**. *Signal Processing* 2001, **18**(4):8–20.
9. Chechetkin VR, Turygin AY: **Size-Dependence of Three-Periodicity and Long-Range Correlations in DNA Sequences**. *Physics Letters A* 1995, **199**:75–80.
10. Chechetkin VR, Turygin AY: **Search of Hidden Periodicities in DNA Sequences**. *Journal of Theoretical Biology* 1995, **175**:477–494.
11. Silverman BD, Linsker R: **A Measure of DNA Periodicity**. *Journal of Theoretical Biology* 1986, **118**(3):295–300.
12. Holste D, et al: **Repeats and Correlations in Human DNA Sequences**. *Physical Review* 2003, **E 67**(06913).
13. Anastassiou D: **Genomic Signal Processing**. *IEEE Signal Processing Magazine* 2001, **18**(4):8–20.
14. Chechetkin VR, Lobzin VV: **Anticodons, Frameshifts, and Hidden Periodicities in tRNA Sequences**. *Journal of Biomolecular Structure and Dynamics* 2006, **24**(2):189–202.
15. Anastassiou D: **Frequency Domain Analysis of Biomolecular Sequences**. *Bioinformatics* 2000, **16**(12):1073–1082.
16. Tiwari S, et al: **Prediction of Probable Genes by Fourier Analysis of Genomic Sequences**. *CABIOS* 1997, **13**:263–270.
17. Akhtar M, Ambikairajah E: **Time and Frequency domain methods for gene and exon prediction in Eukaryotes**. In *Proceedings of ICASSP* 2007:573–576.
18. Voss RF: **Evolution of Long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences**. *Phy. Rev. Lett.* 1992, **68**(25):3805–3808.
19. Fickett JW: **The Gene Identification Problem: An Overview for Developers**. *Computers Chemistry* 1996, **20**:103–118.
20. Bouaynaya N, Schonfeld D: **Non-stationary Analysis of Coding and Non-coding Regions in Nucleotide Sequences**. *IEEE Journal of Selected Topics in Signal Processing* 2008, **2**(3):357–364.
21. Vaidyanathan PP, Yoon BJ: **Gene and Exon Prediction using All Pass-Based Filters**. In *Gensips Proc.* 2003.
22. Vaidyanathan PP, Yoon B: **Digital filter for Gene Prediction Applications**. In *Proc. Asilomar conference* 2003.
23. Akhtar M, Epps J, Ambikairajah E: **On DNA numerical representations for period-3 based exon prediction**. In *Proceedings of the workshop on Genomic Signal Processing and Statistics* 2007.
24. Rushdi A, Tuqan J: **Gene Identification using the Z-curve representation**. *in Proceedings of the* 31^{st} *IEEE ICASSP conference* 2006, **II**:1024–1027.
25. Tuqan J, Rushdi A: **A DSP Approach for Finding the Codon Bias in DNA Sequences**. *IEEE Journal on Selected Topics in Signal Processing* 2008, **2**(3):343–356.

26. Rushdi A, Tuqan J: **An Efficient Algorithm for DNA Discrete Fourier Analysis**. *in Proceedings of the* 3^{rd} *IEEE Cairo International Biomedical Engineering Conference (CIBEC)* 2006, **BI**:1–4.

27. Wang L, Schonfeld D: **Mapping equivalence for symbolic sequences: Theory and applications**. *IEEE Transactions on Signal Processing* 2009, **57**(12):4895–4905.

28. Rushdi A, Tuqan J: **The role of the Symbolic-to-Numerical Mapping in the detection of DNA Periodicities**. In *Proceedings of the workshop on Genomic Signal Processing and Statistics* 2008:1–4.

29. Cristea PD: **Conversion of Nucleotides Sequences into Genomic Signals**. *Journal of Cellular Molecular Medicine* 2002, **6**(2):279–303.

30. Brodzik AK, Peters O: **Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences**. *Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* 2005, **5**:373–376.

31. Coward E: **Equivalence of two Fourier methods for biological sequences**. *journal of mathematical biology* 1997, **36**.

32. Burset M, Guigo R: **Evaluation of Gene Structure Prediction Programs**. *Genomics* 1996, **34**:353–357.

33. Rushdi A, Tuqan J: **Trigonometric Transforms for Finding Repeats in DNA sequences**. In *Proceedings of the workshop on Genomic Signal Processing and Statistics* 2008:1–4.

34. Berger JA, Mitra SK, Astola J: **Power Spectrum Analysis for DNA Sequences**. In *Proc. of the Int. Sym. on Signal Processing and its App.* 2003:29–32.

35. Kotlar D, Lavner Y: **Gene prediction by Spectral Rotation Measure: A new method for Identifying Protein-Coding Regions**. *Genome Research* 2003.

36. Pinto MC: **Finding Periodicities in DNA sequences with a wavelet technique**. *Bioinformatics* 2005.

37. Rushdi A, Tuqan J: **The Filtered Spectral Rotation Measure**. *in Proceedings of the* 40^{th} *IEEE Asilomar Conference on Signals, Systems, and Computers* 2006, :1875–1879.

38. Datta S, Asif A: **A Fast DFT Based Gene Prediction Algorithm for Identification of Protein Coding Regions**. In *Proc. of the ICASSP* 2005:113–116.

$\mathbb{F}$	$\{A, C, G, T\}$, the field of DNA nucleotides
$\mathbb{V}$	$\{0, 1\}$, the field of Voss binary elements
$\mathbb{D}$	A general field of complex valued elements
$\mathbb{F} \longmapsto \mathbb{D}$	Field mapping operation from set $\mathbb{F}$ to set $\mathbb{D}$, resulting in γ sequences $x_l(n)$, where $l = 1, \ldots, \gamma$. For example, when $\mathbb{D} = \mathbb{V}$, $\mathbb{F} \longmapsto \mathbb{D}$ results in $\gamma = 4$ binary sequences, namely: $x_A(n), x_C(n), x_G(n)$, and $x_T(n)$
$x_l(n)$	A discrete time sequence of length N whose elements belong to the mapped field $\mathbb{D}$
$\mathbf{x}_l(n)$	The n^{th} window of length M, extracted from $x_l(n)$, $l = 1, \ldots, \gamma$
$\hat{\mathbf{x}}_l(n)$	The interleaved version of $\mathbf{x}(n)$ with an interleaving factor R, $l = 1, \ldots, \gamma$
$X_l(Rn, \frac{M}{R})$	The ST-DFT of $x_l(n)$, generated using a sliding window of length M and a window shift of length R
$\boldsymbol{\Upsilon}_v(n)$	$[X_A(n)\ X_C(n)\ X_G(n)\ X_T(n)]^T$, the array of the four $\mathbb{V}$-based ST-DFTs
$\boldsymbol{\Upsilon}_d(n)$	$[X_1(n)\ X_2(n) \ldots\ X_\gamma(n)]^T$, the array of the γ $\mathbb{D}$-based ST-DFTs
$X_{lr}(n)$	The r^{th} filtered polyphase component of $X_l(n)$, where $r = 0, 1, \ldots, R-1$ and $l = 1, \ldots, \gamma$
$S_v(n)$	The DNA spectrum computed by adding the magnitude squared of the ST-DFT of the four $\mathbb{V}$-based sequences
$S_d(n)$	The DNA spectrum computed by adding the magnitude squared of the ST-DFT of the γ $\mathbb{D}$-based sequences
$\boldsymbol{\Gamma}_l(n)$	$[X_{l0}(n)\ X_{l1}(n) \ldots X_{l,R-1}(n)]^T$, the array of the R filtered polyphase components $X_{lr}(n)$, $r = 0, 1, \ldots, R-1$ and $l = 1, \ldots, \gamma$
$\mathbf{I}_\gamma$	An identity matrix of size $\gamma \times \gamma$
$\mathbf{C}$	An array of length R whose elements are equally spaced on the unit circle
$\mathbf{h}$	An array of length M/R whose elements are all equal to one
$\mathbf{D}$	$\mathbf{C}^\star\mathbf{C}^T$, an $R \times R$ matrix
$\mathbf{H}$	$\mathbf{I}_R \otimes \mathbf{h}^T$, an $R \times R$ block matrix of $\frac{M}{R} \times 1$ blocks
$\mathbf{W}$	$\mathbf{H}^H\mathbf{D}\mathbf{H}$, an $R \times R$ block matrix of $\frac{M}{R} \times \frac{M}{R}$ blocks
$\mathbf{A}, \mathbf{b}$	The affine transformation matrices of size $\gamma \times 4$ and $\gamma \times 1$ respectively that map the four $\mathbb{V}$-based sequences into the γ $\mathbb{D}$-based sequences.
$\mathbf{B}$	$\mathbf{A}^\mathbf{H}\mathbf{A}$, a 4×4 matrix
$\tilde{\mathbf{C}}$	A complex valued array of R elements
$\tilde{\mathbf{h}}$	A complex valued array of M/R elements
$\tilde{\mathbf{D}}$	$\tilde{\mathbf{C}}^\star\tilde{\mathbf{C}}^T$, an $R \times R$ matrix
$\tilde{\mathbf{H}}$	$\mathbf{I}_R \otimes \tilde{\mathbf{h}}^T$, an $R \times R$ block matrix of $\frac{M}{R} \times 1$ blocks
$\tilde{\mathbf{W}}$	$\tilde{\mathbf{H}}^H\tilde{\mathbf{D}}\tilde{\mathbf{H}}$, an $R \times R$ block matrix of $\frac{M}{R} \times \frac{M}{R}$ blocks

Table 1: Summary of the paper notations

$\{\cdot\}^*$	Matrix complex conjugate
$\{\cdot\}^T$	Matrix transpose
$\{\cdot\}^H$	Matrix hermitian
$\{\otimes\}$	Kronecker product of two matrices
$vec\{.\}$	Vector of columns of a matrix

Table 2: Notation of matrix operations

FIR Window	A_1/A_0	$\Delta\omega$	β	$\Delta\omega_\beta$
Rectangular	-13	$4\pi/(M/R+1)$	0	$1.81\pi R/M$
Bartlett	-25	$8\pi R/M$	1.33	$2.37\pi R/M$
Hanning	-31	$8\pi R/M$	3.86	$5.01\pi R/M$
Hamming	-41	$8\pi R/M$	4.86	$6.27\pi R/M$
Blackman	-57	$12\pi R/M$	7.04	$9.19\pi R/M$

Table 3: FIR window Specifications: relative peak side lobe A_1/A_0 in dB, approximate width of main lobe $\Delta\omega$, equivalent Kaiser window coefficient β, and transition width $\Delta\omega_\beta$

Transform	α	a	b	θ_r
ST-DFT	1	1	0	$-2\pi r/R$
ST-DCT	-1	$1/2$	$-1/2$	$(2r+1)\pi/2R$
ST-DST	-1	$1/2j$	$-1/2j$	$(2r+1)\pi/2R$
ST-DHT	1	$\frac{1}{2}(1-j)$	$-\frac{1}{2}(1-j)$	$2\pi r/R$

Table 4: Parameter settings in Figure 15 to compute the short time Fourier, cosine, sine, and Hartley transforms.

ST-DFT	$\tilde{\mathbf{h}} = \mathbf{h} = \{(1)^i, i = 1, 2, \ldots M/R\}$ $\tilde{\mathbf{C}} = \mathbf{C} = \{e^{-j2\pi r/R}, r = 1, 2, \ldots R\}$
ST-DCT	$\tilde{\mathbf{h}} = \{(-1)^i, i = 1, 2, \ldots M/R\}$ $\tilde{\mathbf{C}} = \{\cos((2r+1)\pi/2R), r = 1, 2, \ldots R\}$
ST-DST	$\tilde{\mathbf{h}} = \{(-1)^i, i = 1, 2, \ldots M/R\}$ $\tilde{\mathbf{C}} = \{\sin((2r+1)\pi/2R), r = 1, 2, \ldots R\}$
ST-DHT	$\tilde{\mathbf{h}} = \mathbf{h} = \{(1)^i, i = 1, 2, \ldots M/R\}$ $\tilde{\mathbf{C}} = \{\cos(2\pi r/R) + \sin(2\pi r/R), r = 1, 2, \ldots R\}$

Table 5: Modified arrays $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{C}}$ to compute the short time Fourier-, cosine-, sine-, and Hartley-based DNA spectrum of (38).

$\hat{\mathbf{x}}(n)$, $\mathbf{W}$	Real multiplications	Real additions
real,real	$M(M+1)$	M^2-1
real,complex	$2M(M+1)$	$2(M^2-1)$
complex,real	$2M(M+1)$	$2(M^2-1)$
complex,complex	$4M(M+1)$	$2(2M^2+M-1)$

Table 6: Real multiplications and additions needed for the evaluation of (39) and (15).