Periodic power spectrum with applications in detection of latent periodicities in DNA sequences

Changchuan Yin1 · Jiasong Wang2

Received: 21 May 2015 / Revised: 19 February 2016 / Published online: 4 March 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Periodic elements play important roles in genomic structures and functions, yet some complex periodic elements in genomes are difficult to detect by conventional methods such as digital signal processing and statistical analysis. We propose a periodic power spectrum (PPS) method for analyzing periodicities of DNA sequences. The PPS method employs periodic nucleotide distributions of DNA sequences and directly calculates power spectra at specific periodicities. The magnitude of a PPS reflects the strength of a signal on periodic positions. In comparison with Fourier transform, the PPS method avoids spectral leakage, and reduces background noise that appears high in Fourier power spectrum. Thus, the PPS method can effectively capture hidden periodicities in DNA sequences. Using a sliding window approach, the PPS method can precisely locate periodic regions in DNA sequences. We apply the PPS method for detection of hidden periodicities in different genome elements, including exons, microsatellite DNA sequences, and whole genomes. The results show that the PPS method can minimize the impact of spectral leakage and thus capture true hidden periodicities in genomes. In addition, performance tests indicate that the PPS method is more effective and efficient than a fast Fourier transform. The computational complexity of the PPS algorithm is $O(N)$. Therefore, the PPS method may have a broad range of applications in genomic analysis. The MATLAB programs for implementing the PPS method are available from MATLAB Central (http://www.mathworks.com/matlabcentral/fileexchange/55298).

Changchuan Yin
cyin1@uic.edu

1 Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7045, USA
2 Department of Mathematics, Nanjing University, Nanjing, Jiangsu 210093, China
Keywords Fourier transform · Periodic power spectrum · Spectral leakage · DNA genome · Periodicity

Mathematics Subject Classification 42A38 · 92D20

1 Introduction

Regions of inexact repeats are abundant in genomes from bacteria to mammals. Repetitive elements comprise over two-thirds of the human genome (Koning et al. 2011). In bacteria, repetitive sequences may account for 5–10% of a total genome. Repetitive elements are essential for genomic structures and functions (Shapiro and Sternberg 2005). A 3-periodicity of a DNA sequence indicates protein coding regions (Silverman and Linsker 1986; Yin and Yau 2005). Pervasive hidden 10–11 bp periodicities in complete genomes reflect DNA folding structure (Herzel et al. 1999). In bacterial genomes, repetitive DNA sequences form alternative conformations and thus cause genetic instability (Wojcik et al. 2012). Conversely, in eukaryotic genomes, repetitive DNA sequence elements are also fundamental to cooperative molecular interactions forming nucleoprotein complexes (Shapiro and Sternberg 2005). The 10 bp and other longer periodicities in genomes may associate with nucleosome organization (Segal et al. 2006; Chen et al. 2008). An expanded simple repeat accounts for human neurodegenerative diseases (Sutherland and Richards 1995; Renton et al. 2011). Short tandem repeats (STRs) have a wide range of applications, including medical genetics, forensics, and genetic genealogy (Gymrek et al. 2012). In addition, repeats in genomes often cause problems for assembling, and detecting insertion or deletion in next-generation sequencing (Treangen and Salzberg 2011; Wojcik et al. 2012).

Repetitive sequences and periodicities in genomes vary in size, scale and complexity (Trifonov 1998; Hauth and Joseph 2002). The sizes of repeats may range from two base pairs to hundreds of base pairs. Repetitive sequences may disperse widely in a relatively long sequence range or exist as tandem arrays of simple sequence composition. Some repeats contain partial periodic sequence patterns; other repeats have hidden periodicities due to multiple periodic components (Korotkov et al. 2003). Therefore, partial and hidden repeat signals in DNA sequences are difficult to analyze through straightforward observation and sequence comparison.

The hidden periodicities and repetitive features of DNA and protein sequences are primarily studied under digital signal processing approaches such as Fourier Transform (Sharma et al. 2004; Buchner and Janjarasjitt 2003), Ramanujan–Fourier Transform (Yin et al. 2014b), and Wavelet transform (Murray et al. 2002; Wang and Stein 2010). DNA sequences are first converted to numerical sequences and then undergo Fourier transform for power spectrum analysis in the sequences (Afreixo et al. 2004). Other important periodicity detection methods include statistical analysis (Epps et al. 2011), maximum likelihood estimation (de Koning et al. 2011), information decomposition (Korotkov et al. 2003), direct frequency mapping (Glunčić and Paar 2013), and chaos game representation (CGR) (Messaoudi et al. 2014), etc. For a review of these methods, one may refer to Grover et al. (2012). Despite many studies on hidden periodicities in DNA sequences, the complexity of repetitive sequences has made it difficult to
reduce the high background noise in Fourier analysis and statistical methods. Fast and accurate identification of all hidden periodicities in DNA sequences still poses a challenge in genomic analysis (Suvorova et al. 2014; Epps et al. 2011; Illingworth et al. 2008).

The Discrete Fourier transform (DFT) is widely used technique in identifying periodic signals in DNA or protein sequences (Anastassiou 2000). Mapping DNA or protein sequences into the Fourier frequency domain is an effective way of processing the data so that their global periodic characteristics can be revealed. The Fourier transform has been extensively used in different bioinformatics researches. It is used in genomic period analysis of DNA and protein sequences (Sharma et al. 2004; Chechetkin and Turygin 1995), protein docking (Ritchie and Kemp 2000), protein coding region prediction (Tiwari et al. 1997; Yin and Yau 2007), similarity analysis of DNA sequences (Yin et al. 2014a; Yin and Yau 2015; Hoang et al. 2015), protein docking (Katchalski-Katzir et al. 1992), etc. Despite widespread applications, the Fourier transform still has limitations, particularly a need for high computational resources (Scargle 1982). DFT is linear and assumes stationarity in data. It works best with relatively long datasets. If data do not meet these assumptions, the Fourier Transform provides less than adequate results and cannot precisely detect certain periodic signals. For example, DFT is ineffective in detecting 3-periodicity of short exon sequences (Kotlar and Lavner 2003). Furthermore, the DFT spectral analysis of biological sequences is often impacted by spectral leakage and high level of background noise (Datta and Asif 2005). These limitations of DFT highlight the need of effective techniques for detecting periodicities of biological sequences.

To overcome the shortcomings in the Fourier analysis of DNA sequences, in this paper, we present a novel algorithm for computing periodic power spectra (PPS) based on nucleotide distribution of DNA sequences. As one of the applications, we apply the PPS algorithm to the identification of hidden periodicities in DNA sequences. Experimental results demonstrate that this algorithm is effective and efficient in quantitatively capturing hidden periodicities in DNA sequences and outperforms the Fourier transform and statistical methods.

2 Methods and algorithms

2.1 Numerical representations of DNA sequences

DNA molecules consist of four linearly joined nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G). A DNA sequence can be represented as a permutation of four characters A, T, C, G of different lengths. To apply digital signal processing in DNA sequence analysis, the character strings of DNA molecules must be mapped into one or more numerical sequences (Wang and Johnson 2002; Yin 2015). The numerical representation of DNA symbolic sequence is subject to digital signal processing in order to extract information that corresponds to biological functions. One commonly used mapping methods is to use 4 dimensional (4D) binary indicator sequences (Voss 1992). In the 4D indicator mapping method, a DNA sequence of length \( N \), denoted as, \( s(0), s(1), \ldots, s(N-1) \), is decomposed into four binary
Table 1  The Voss 4D binary indicator mappings of a short DNA sequence

| DNA      | T | A | G | C | C | T | G | C | T | G | A | T |
|----------|---|---|---|---|---|---|---|---|---|---|---|---|
| $u_A$    | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |   |
| $u_T$    | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |   |
| $u_C$    | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |   |
| $u_G$    | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |   |

indicator sequences, $u_A(n)$, $u_T(n)$, $u_C(n)$, and $u_G(n)$, which indicate the presence or absence of four nucleotides, $A$, $T$, $C$, and $G$ at the $n$-th position, respectively. The 4D indicator mapping of DNA sequences is defined as follows:

$$u_\alpha(n) = \begin{cases} 1, & s(n) = \alpha \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (1)$$

where $\alpha \in A, C, G, T$, $n = 0, 1, 2, \ldots, N - 1$. The four indicator sequences correspond to the appearance of the four nucleotides $A$, $T$, $C$, and $G$ at each position of DNA sequences. For example, an indicator sequence, $u_A(n) = 0001010111 \ldots$, indicates that nucleotide $A$ presents in positions of 4, 6, 8, 9, and 10 of a DNA sequence. An example illustrates the Voss 4D indicator mapping of an example DNA sequence (Table 1).

2.2 Fourier power spectrum

The Fourier transform is a transformation of signal in time domain to new values in frequency domain (Welch 1967; Anastassiou 2000). It gives a unique representation of the original signal in frequency domain and can reveal the global periodic characteristics of a signal in frequency domain. The Fourier transform attempts to elucidate a time series as a weighted sum of sinusoidal basis elements, where the time series can be closely approximated by such elements. Let $X$ be the DFT of real number time series $x$ of length $N$. Then $X(k)$ is defined as

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-i2\pi \frac{k}{N} n}, k = 0, 1, \ldots, N - 1$$  \hspace{1cm} (2)$$

where $i = \sqrt{-1}$. The frequency domain vector $X$ contains all information about the original signal $x$, and can recover the signal. The DFT power spectrum of the signal $x(n)$ at the frequency $k$ is defined as

$$PS(k) = X(k)X(k)^*, k = 0, 1, \ldots, N - 1$$  \hspace{1cm} (3)$$

where $X(k)^*$ denotes the complex conjugate of $X(k)$. The Fourier power spectrum $PS(k)$ reflects the periodic features of the signal $x$ in frequency domain.
It should be noted that the discrete nature of the DFT sometimes may yield intrinsic quantization errors and sampling artifacts. For periodic sinusoidal signal, \( A = \sin(2\pi fnT) \), where \( N \) is the number of samples and \( T \) is the distance between two samples, and \( n = 0, 1, 2, \ldots N - 1 \). The Fourier power spectrum of the signal has a peak proportional to the amplitude at frequency \( f \). If the frequency \( f \) of a signal is not a multiple of \( F = 1/NT \), a distorted DFT is obtained. This phenomenon is referred to spectral leakage (Costa and Melucci 2010; Lyon 2009). For instance, if the length of a signal is an odd number, all periodic components in the signal may show some level of leakage in the Fourier power spectrum. Due to a spectral leakage, recognizing correct frequencies of a signal becomes difficult. Spectral leakage should thus be avoided in digital signal processing.

The DFT spectral analysis of DNA sequences may detect any latent or hidden periodic signal in the original sequences. It may discover approximate repeats that are difficult to detect by tandem repeat search. To apply Fourier transform to a DNA sequence, the DNA sequence is first converted to four binary indicator sequences. Let \( U_\alpha \) and \( PS_\alpha \) be the Fourier transform and corresponding Fourier power spectrum of the binary indicator sequence \( u_\alpha, \alpha \in \{A, T, C, G\} \) of the sequence, respectively. Then, the Fourier power spectrum of the DNA sequence is the sum of Fourier power spectra \( PS_\alpha \).

\[
PS(k) = \sum_{\alpha \in \{A, T, C, G\}} PS_\alpha(k).
\]  

The Fourier power spectrum of DNA sequences can be used to identify periodicities in the sequences. For example, most of protein coding regions show a prominent peak at frequency \( k = N/3 \) in Fourier power spectrum because of non-uniform distributions of nucleotides in the three codon positions (Yin and Yau 2005). The Fourier power spectrum is large when the nucleotide has a significant tendency to appear at about every \( p \) positions. In particular, when \( k = N/3 \), one the nucleotide \( \alpha \in \{A, T, C, G\} \) tends to appear at a certain codon position. Because the Fourier power spectra of real number series are symmetric, plotting Fourier power spectrum in this paper only shows the first half of a power spectrum. Furthermore, DFT power spectrum at frequency zero is a constant and equivalent to the sum of data points, thus the power spectrum at frequency zero is not included in the figures of this paper. For the comparison of the PPS method and DFT, the fast Fourier transform (FFT) function in MATLAB is used in this study, which does not require the length of signal be power of 2.

To avoid the spectral leakage in Fourier transform at a specific periodicity in DNA sequences, we pad zeros to extend the lengths of 4D indicator vectors of DNA sequences to a multiple of periodicity length. For example, to compute accurate power spectrum at periodicity 3, the length of a DNA sequence after padding zeros shall be a multiple of 3. However, padding zeros to a certain length only mitigates the problem for revealing a specific periodicity, not all periodicities.

### 2.3 Periodic power spectrum

The strength of a periodicity within a 1D real number signal is determined by the distribution of signal values on periodic positions. The distribution of signal values
on periodic positions can be measured by a congruence derivative vector of a signal, which was introduced by Wang et al. (2012). The congruent derivative vector of a 1D real number signal is defined as follows.

**Definition 1** For a real number signal \( x \) of length \( N \), the element of congruent derivative vector \( f_p \) of the signal \( x \), for periodicity \( p \), is defined as

\[
f_p(q) = \sum_{\text{mod}(n,p) = q} x(n)
\]

where \( q = 0, 1, \ldots, p - 1 \), \( n = 0, 1, \ldots, N - 1 \)

where \( \text{mod}(n, p) \) is the modulo operation and returns the remainder after division of \( n \) by \( p \), then \( f_p = (f_p(0), f_p(1), \ldots, f_p(p - 1)) \).

Correspondingly, for DNA sequences, it has been shown that Fourier power spectrum at a periodicity \( p \) is determined by a periodic nucleotide distribution of the sequences. For instance, the 3-base periodicity is determined by a non-uniform distribution of nucleotides on three codon positions (Yin and Yau 2005, 2007). Here, we extend the congruent derivative vectors from 1D real signals to DNA sequences so that they reflect the periodic nucleotide distributions of the sequences. For example, a DNA sequence, AGTTAACGCCTAGCC, is first converted into 4D Voss binary indicator sequences, \( u_A(n) \), \( u_T(n) \), \( u_C(n) \), and \( u_G(n) \). Each indicator sequence is considered as a 1D real signal. The congruence vectors for periodicity 3 of the indicator sequences are:

\[
\begin{align*}
    f_A &= [1, 1, 2] \\
    f_T &= [1, 1, 1] \\
    f_C &= [2, 1, 2] \\
    f_G &= [1, 2, 0]
\end{align*}
\]

We note that these congruence derivative vectors describes nucleotide distributions on periodic-3 positions in the sequence.

We may consider a transform in frequency domain of 1D real number signal \( x \) as the projection of the signal onto some periodic shift sequences. Let \( \omega_p = e^{-i \frac{2\pi}{p}} \) be the \( p \)-th root of unity, then we have the following periodic shift sequences as basis vectors.

\[
\delta_p^q(n) = \begin{cases} 
\omega_p^q & \text{if mod}(n - q, p) = 0 \\
0 & \text{otherwise}
\end{cases}
\]

where \( q = 0, 1, \ldots, p - 1 \)

Table 2 is an example of periodic shift sequences \( \delta_p^q \) as basis vectors for periodicity 4.

We propose here a periodic transform of a real number signal \( x \) at periodicity \( p \) as the projection a signal onto the corresponding periodic basis vectors \( \delta_p^q \) that is described in Eq. (7). The periodic transform is defined as

| \( n \) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | ... |
|--------|---|---|---|---|---|---|---|---|-----|
| \( \delta_4^0(n) \) | \( \omega_4^0 \) | 0 | 0 | 0 | \( \omega_4^0 \) | 0 | 0 | 0 | ... |
| \( \delta_4^1(n) \) | 0 | \( \omega_4^1 \) | 0 | 0 | 0 | \( \omega_4^1 \) | 0 | 0 | ... |
| \( \delta_4^2(n) \) | 0 | 0 | \( \omega_4^2 \) | 0 | 0 | 0 | \( \omega_4^2 \) | 0 | ... |
| \( \delta_4^3(n) \) | 0 | 0 | 0 | \( \omega_4^3 \) | 0 | 0 | 0 | \( \omega_4^3 \) | ... |
Definition 2

\[ XP(p) = \sum_{n=0}^{N-1} \sum_{q=0}^{p-1} x(n) \delta_p^q(n), \quad p = 1, 2, \ldots, N. \]

Specially, if the length of a signal is equal to a multiple of the periodicity \( p \), we have the following theorem.

**Theorem 1** If \( N \) is equal to a multiple of the periodicity \( p \), the projection of a signal \( x \) onto periodic basis sequences \( \delta_p^q \) is equivalent to the Fourier transform of the congruent derivative vector of the signal.

The proof of Theorem 1 is provided by Wang et al. (2012). Let \( \delta_p = (\omega_0^p, \omega_1^p, \ldots, \omega_{p-1}^p) \); the periodic transform at periodicity \( p \) may be re-written as

\[ XP(p) = f_p \delta_p^T \]

where \( \delta_p^T \) is the transpose of \( \delta_p \).

The power spectrum at periodicity \( p \) is then expressed as

\[ PPS(p) = f_p (\delta_p^T \delta_p) f_p^T \]

where \( f_p^T \) is the transpose of \( f_p \) and \( \delta_p^T \) is the transpose of \( \delta_p \). The matrix formed by \( \delta_p^T \delta_p \) only depends on periodicity \( p \) and is a complex symmetric matrix, whose main diagonal elements are one. If we add the corresponding upper and lower entries of the matrix, we then get a lower-triangular real number matrix, \( Sp \), called spectral transform matrix \( Sp \), which can be used to compute the periodic power spectrum at periodicity \( p \). Spectral transformation matrix \( Sp \) is defined as:

\[
S_p(k, j) = \begin{cases} 
2 \cos \frac{2\pi(k-1)}{p} \cos \frac{2\pi(j-1)}{p} + 2 \sin \frac{2\pi(k-1)}{p} \sin \frac{2\pi(j-1)}{p}, & \text{if } k > j \\
1, & \text{if } k = j \\
0, & \text{if } k < j 
\end{cases}
\]

Based on theories and equations above, we propose the following algorithm (Algorithm 1) to compute the power spectrum of a 1D real number signal using congruence derivative vector and spectral transform matrix \( Sp \). Because the power spectrum from our algorithm corresponds to a specific periodicity, while the Fourier power spectrum corresponds to specific frequencies, we name the proposed method as PPS (periodic power spectrum) algorithm.
2.4 Algorithm for computing periodic power spectrum of a DNA sequence

From the definition of DFT power spectrum and Theorem 1, we can get the power spectrum of a DNA sequence after converting the sequence into 4D binary indicator sequences. For example, the power spectrum to describe 3-base periodicity property of protein coding regions in a DNA sequence can be obtained by its four congruence derivative vectors of the DNA sequence without Fourier transform (Yin and Yau 2007). Here, using PPS for 1D real signal, we propose the following algorithm (Algorithm 2) to compute the power spectrum at any periodicities in a DNA sequence.

**Algorithm 1:** Algorithm for computing PPS at periodicity $p$ of a real number signal.

**Algorithm 2:**

1. **Input:** a 1D real number signal $x$ of length $N$, periodicity $p$
2. **Output:** PPS at periodicity $p$
3. **Step:**
   1. Generate a vector $C = [1, \cos(2\pi/p), \cos(4\pi/p), \ldots, \cos(2(p-1)\pi/p)]$.
   2. Generate a vector $V = [0, \sin(2\pi/p), \sin(4\pi/p), \ldots, \sin(2(p-1)\pi/p)]$.
   3. Obtain a matrix $U = C^TC + V^TV$, where $C^T$ and $V^T$ are the transposes of vectors $C$ and $V$, respectively.
   4. Construct the spectrum transform matrix $S_p$ of size $p \times p$ from the matrix $U$:
      - If $k > j$ then return $S_p(k, j) = U(k, j) + U(j, k)$
      - Else return $S_p(k, j) = 0$
   5. Compute the congruent derivative vector $f_p$ of signal at periodicity $p$ (Equation (5)).
   6. Compute the periodic power spectrum $PPS(p)$ by $S_p$ and $f_p$:
      $$PPS(p) = f_p^T S_p f_p^T$$
      where $f_p^T$ is the transpose of congruent derivative vector $f_p$.

2.4 Algorithm for computing periodic power spectrum of a DNA sequence

After we obtain the spectrum transform matrix for periodicity $p$, $S_p$, and congruent derivative vectors from the Algorithm 2 steps 1–3, we can compute the power spectrum of DNA sequences for periodicity $p$. Let $f_{a0}, f_{a1}, \ldots, f_{ap-1}, \alpha \in \{A, T, C, G\}$ be the elements of a congruent derivative vector, $f_{\alpha}$, of periodicity $p$ in a DNA sequence, which are the occurring frequency of nucleotide $\alpha$, the Algorithm 2 can be used to compute the $PPS(p)$ at periodicity $p$ of the DNA sequence. Note that $PPS(p)$ can also be computed as follows:
**Input:** a DNA sequence of length \( N \), periodicity \( p \)

**Output:** PPS at periodicity \( p \)

**Step:**
1. Convert the DNA sequence into four binary indicator sequences \( u_\alpha, \alpha \in \{A, T, C, G\} \).
2. Compute the congruent derivative vectors \( f_\alpha \) of the sequences \( u_\alpha, \alpha \in \{A, T, C, G\} \), at periodicity \( p \).
3. Compute the spectrum transform matrix \( S_p \) of size \( p \) (refer to algorithm 1, steps 1-4).
4. Compute the power spectrum \( PPS(p) \) by the four congruent derivative vectors \( f_\alpha \) of periodicity \( p \):

\[
PPS(p) = \sum_{\alpha \in \{A, T, C, G\}} f_\alpha \ast S_p \ast f_\alpha^T
\]

where \( f_\alpha^T \) is the transpose of congruent derivative vector \( f_\alpha \).

**Algorithm 2:** Algorithm for computing PPS at periodicity \( p \) of a DNA sequence.

\[
PPS(p) = \sum_{\alpha \in \{A, T, C, G\}} \left( \sum_{q=0}^{p-1} f_{\alpha q}^2 + \sum_{k=0, j=0, k > j}^{p-1} S_p(k, j) f_{\alpha j} f_{\alpha k} \right)
\]

Because short periodicities often receive much attention in genome study, from Algorithm 2, we provide the following formulas for calculating the PPS spectrum of short periodicities in DNA sequences. We first construct the spectrum transform matrices for special short periodicities, \( S_2, S_3, S_4, S_5, S_6 \), as shown in Table 3, then we can compute the PPS spectrum at a few short periodicities using these spectral transform matrices.

\[
PPS(2) = \sum_{\alpha \in \{A, T, C, G\}} \left( f_{\alpha 0}^2 + f_{\alpha 1}^2 - 2f_{\alpha 0} f_{\alpha 1} \right)
\]

\[
PPS(3) = \sum_{\alpha \in \{A, T, C, G\}} \left( \sum_{q=0}^{2} f_{\alpha q}^2 - f_{\alpha 0} f_{\alpha 1} - f_{\alpha 0} f_{\alpha 2} - f_{\alpha 1} f_{\alpha 2} \right)
\]

\[
PPS(4) = \sum_{\alpha \in \{A, T, C, G\}} \left( \sum_{q=0}^{3} f_{\alpha q}^2 - 2f_{\alpha 0} f_{\alpha 2} - 2f_{\alpha 1} f_{\alpha 1} \right)
\]

\[
PPS(5) = \sum_{\alpha \in \{A, T, C, G\}} \left( \sum_{q=0}^{4} f_{\alpha q}^2 \begin{array}{c} + \frac{-1 + \sqrt{5}}{2} (f_{\alpha 0} f_{\alpha 1} + f_{\alpha 1} f_{\alpha 2} + f_{\alpha 2} f_{\alpha 3} + f_{\alpha 3} f_{\alpha 4} + f_{\alpha 4} f_{\alpha 0}) \\ + \frac{-1 - \sqrt{5}}{2} (f_{\alpha 0} f_{\alpha 2} + f_{\alpha 1} f_{\alpha 3} + f_{\alpha 2} f_{\alpha 4} + f_{\alpha 3} f_{\alpha 5} + f_{\alpha 4} f_{\alpha 5}) \end{array} \right)
\]

\[
PPS(6) = \sum_{\alpha \in \{A, T, C, G\}} \left( \sum_{q=0}^{5} f_{\alpha q}^2 \begin{array}{c} + (f_{\alpha 0} f_{\alpha 1} + f_{\alpha 1} f_{\alpha 2} + f_{\alpha 2} f_{\alpha 3} + f_{\alpha 3} f_{\alpha 4} + f_{\alpha 4} f_{\alpha 5} + f_{\alpha 5} f_{\alpha 0}) \\ - (f_{\alpha 0} f_{\alpha 2} + f_{\alpha 0} f_{\alpha 4} + f_{\alpha 1} f_{\alpha 3} + f_{\alpha 3} f_{\alpha 5} + f_{\alpha 4} f_{\alpha 5} + f_{\alpha 2} f_{\alpha 4}) \\ - 2(f_{\alpha 0} f_{\alpha 3} + f_{\alpha 1} f_{\alpha 4} + f_{\alpha 2} f_{\alpha 5}) \end{array} \right)
\]
Table 3  Spectrum transform matrices of some short periodicities

| Periodicity | Spectrum transform matrix $S_p$ |
|-------------|----------------------------------|
| $P_2$       | $\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}$ |
| $P_3$       | $\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix}$ |
| $P_4$       | $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & -2 & 0 & 1 \end{bmatrix}$ |
| $P_5$       | $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{\sqrt{5} - 1}{2} & 1 & 0 & 0 & 0 \\ -\frac{\sqrt{5} + 1}{2} & \frac{\sqrt{5} - 1}{2} & 1 & 0 & 0 \\ -\frac{\sqrt{5} + 1}{2} & -\frac{\sqrt{5} + 1}{2} & \frac{\sqrt{5} - 1}{2} & 1 & 0 \\ \frac{\sqrt{5} - 1}{2} & -\frac{\sqrt{5} + 1}{2} & -\frac{\sqrt{5} + 1}{2} & \frac{\sqrt{5} - 1}{2} & 1 \end{bmatrix}$ |
| $P_6$       | $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ -2 & -1 & 1 & 1 & 0 & 0 \\ -1 & -2 & -1 & 1 & 1 & 0 \\ 1 & -1 & -2 & -1 & 1 & 1 \end{bmatrix}$ |

To compare the strengths of periodic signals in DNA sequences of different lengths, we use the PPS signal-to-noise ratio (SNR) at a periodicity $p$ of a DNA sequence. SNR at periodicity $p$ is defined as the magnitude of the PPS power spectrum divided by average DFT power spectrum. From our previous study, the average DFT power spectrum of a DNA sequence of length $N$ is equal to $N$ (Yin and Yau 2007). Therefore, the SNR of PPS spectrum of a DNA sequence at periodicity $p$ is defined as the PPS power spectrum at periodicity $p$ divided by the length of the sequence. In this paper, we chose 1 as the SNR threshold to distinguish a true periodicity and background noise.

$$SNR(p) = \frac{PPS(p)}{N}.$$  \hspace{1cm} (16)

It is worth noting that the PPS method directly computes the periodic power spectrum from congruent derivative vector, and does have sampling artifacts, thus, the spectral leakage does not exist in the PPS method.

3 Results

3.1 PPS power spectrum analysis of a periodic signal

To illustrate the effectiveness of the PPS method for identifying hidden periodicities in digital signals, we computed the PPS spectrum of the following periodic sinusoidal signal. The sinusoidal signal of length $N$, $N = 300$, consists of sine and cosine signals
with periodicities 20 and 50, and is corrupted by white Gaussian noise.

\[ x(n) = \sin \left( 2\pi \frac{n}{20} + \frac{\pi}{4} \right) + \cos \left( 2\pi \frac{n}{50} + \frac{\pi}{4} \right) + \text{noise} \]

Due to the impact of white Gaussian noise, the two periodicities 20 and 50 in the original signal are buried and hidden by noise (Fig. 1a). However, Fourier power spectrum of the signal can reveal the two periodicities (Fig. 1b). Since the Fourier power spectrum plot is over frequency domain, high peaks are at frequencies \( f = 15 \) and 6 (\( f = N/p, p = 20, 50 \)) in the Fourier spectrum, corresponding to the periodicities 20 and 50, respectively. The PPS spectrum of the signal can clearly discover two pronounced peaks at positions 20 and 50 (Fig. 1c), corresponding to two periodicities 20 and 50 in the original signal. The magnitudes of periodicities 20 and 50 of the PPS spectrum and Fourier spectrum are the same. These results verify that the PPS spectrum is equal to the DFT spectrum if there is no spectrum leakage, i.e., the length of a signal is a multiple of periodicities. In comparison with Fourier transform, the PPS spectrum directly identifies periodicities in a signal from periodicity positions, while Fourier power spectrum indirectly identifies periodicities from frequency positions. In addition, the results in Fig. 1b, c show that the PPS spectrum has less background noise than the Fourier power spectrum. The putative reason for reduced noise in the PPS spectrum is due to no spectral leakage.

It is worth noting that the PPS spectrum of the sinusoidal signal shows a smooth and global trend when periodicity lengths are increased (Fig. 1c). The Fourier spectra of long periodicities represent low frequency features of a signal. The low frequency components of Fourier power spectrum usually demonstrate the fundamental features of a signal (Agrawal et al. 1993; Foster 1995). Therefore, the pattern of PPS spectrum of long periodicities may represent the intrinsic characters and trends of a signal. These results demonstrate the advantages of the PPS method in periodicity analysis on real number signals.

3.2 PPS power spectrum analysis of DNA sequences

We investigated the impacts of spectral leakages on Fourier power spectrum and the PPS spectrum by computational simulations. To this end, we generated an artificial DNA sequence with periodicity 5, denoted as N130P5 (N130P5:CCATATCCGATCGCAGCGGTGCTTTTTATCTCAGCTATCGAATGGGCTCGAGGACCGCGGCTGTCTATAGAAAAATTATAAATGATATTGATCCGAGTAGGGTCCCACTCGGTGGCGGGGCACTTCAAA). The DNA sequence N130P5 contains 6 copies of imperfect 5-base repeat ATCGA. A deletion mutant of the DNA sequence is constructed by deleting two nucleotides AA at the 3\(^\prime\) end, denoted as N130P5-D2 (N130P5-D2:CCATATCCGATCGCAGCGGTGCTTTTTATCGCTATCGAATGGGCTCGAGGACCGCGGCTGTCTATAGAAAAATTATAAATGATATTGATCCGAGTAGGGTCCCACTCGGTGGCGGGGCACTTCAAAATTATAAATGATATTGATCCGAGTAGGGTCCCACTCGGTGGCGGGGCACTTCAAA). The Fourier power spectrum and the PPS spectrum of the DNA sequences are shown in Fig. 2 and Table 4.
Fig. 1 Power spectrum analysis of sinusoidal signal collapsed with white noise. a Original sinusoidal signal. b DFT power spectrum of the signal. c PPS power spectrum of the signal.
Fig. 2  Power spectrum analysis of artificial DNA sequences with periodicity 5. a Fourier power spectra of DNA sequence N130P5 and N130P5-D2 with zero padding. b Fourier power spectrum of N130P5-D2 without zero padding. c PPS spectra of and N130P5 and N130P5-D2 with or without zero padding.
Table 4  Comparison of Fourier power spectrum and PPS spectrum on artificial DNA sequences N130P5 and N130P5-D2 with periodicity 5

| Sequence                  | DFT: $P S\left(\frac{N}{5}\right)$ | PPS: $P S(5)$ |
|---------------------------|-------------------------------------|---------------|
| N130P5                    | 361.9837                            | 361.9837      |
| N130P5-D2 (with padding zeros) | 335.8034                            | 335.8034      |
| N130P5-D2 (no padding zeros) | 212.0118                            | 335.8034      |

The Fourier power spectrum and the PPS spectrum at the periodicity 5 are the same, the peak value is 361.9837 (Fig. 2a, c). The results verify that the PPS spectrum and Fourier spectrum at periodicity 5 of the DNA sequence are the same when the length of a DNA sequence is a multiple of periodicity 5. The results validate the accuracy of the PPS algorithm in computing power spectrum.

If the deletion mutant N130P5-D2 is padded with two zeros, Fourier power spectra of N130P5-D2 with zero padding and N130P5 are very similar (Fig. 2a). If there is no zero padding in the N130P5-D2 sequence (Fig. 2b), the Fourier power spectrum at periodicity 5 becomes 212.0118 and is much different from the corresponding values from N130P5-D2 with padding zeros (335.803) or N130P5 (335.8034). The reason for these differences is that the length of the deletion mutant is not a multiple of periodicity 5. The periodicity 5 of sequence N130P5-D2 is reduced and hidden in Fourier spectrum because of spectral leakage (Fig. 2b). Padding two zeros in N130P5-D2 to extend the length to a multiple of periodicity 5 can recover the spectral leakage in Fourier power spectrum (Fig. 2a, b). However, the PPS spectra of N130P5 and N130P5-D2 with or without padding zeros are similar (Fig. 2c). This is because there are only two nucleotide deletions in N130P5-D2 and the PPS method is not affected by the spectral leakage. The results from PPS spectra agree with theoretical analysis (Eq. 7). Hence the PPS spectrum can be considered as zero padding for each periodicity. These results indicate that Fourier spectrum may not identify some hidden periodicities due to spectrum leakage, but the PPS spectrum can detect these hidden periodicity in DNA sequences of any lengths.

With additional tests on DNA sequences of different lengths, we confirm that the PPS spectrum and Fourier power spectrum are the same for periodicity $p$ only when the length of a DNA sequence is equal to a multiple of $p$. If the length of the DNA sequence is not a multiple of the periodicity $p$, the Fourier power spectrum is not equivalent to the PPS spectrum. A periodic signal may be hidden by other periodicities in Fourier power spectrum analysis. Thus the Fourier power spectrum of a DNA sequence is affected by the length of the sequence, but the PPS spectrum can overcome the length problem as in Fourier spectrum analysis. This is a special advantage of the PPS method.

We further assessed the PPS spectrum on the well-studied 3-periodicity of exon sequence. The test sequence is the last exon of F56F11.4 gene in *C. elegans* (GenBank: *FO081497, 14275 to 14625 region, 351 bp*). The F56F11.4 gene has been extensively studied and shows 3-periodicity in its exons (*Anastassiou 2000; Wang and Stein 2010*). Both the Fourier power spectrum and the PPS spectrum display a significant peak for periodicity 3 (Fig. 3a, b), but additional periodicities 7 and 14
Fig. 3 Power spectra of the last exon of F56F11.4 gene in *C. elegans*. \(a\) Fourier power spectrum. \(b\) PPS spectrum. \(c\) PPS spectra of DNA walk. \(d\) PPS spectrum SNRs of sliding windowed sequence for different periodicities, window size = 60 bp.
are detected in the PPS spectrum. We then examined these two periodicities using the DNA walk and sliding window approaches as described in our previous study (Yin and Yau 2007). Figure 3b is the PPS spectra of periodicities 3 and 7 using DNA walk. The strengths of the PPS spectrum for the two periodic signals increase when the length of DNA walk increases. This result confirms presence of two periodic signals. From the sliding window plots, we may locate the approximate regions in the gene for these two periodic signals as shown in Fig. 3c. Since the periodicities 7 and 14 can only be identified in PPS spectrum, but are difficult to be identified by Fourier power spectrum (Fig. 3a), the result indicates that the PPS method can capture more periodicities than Fourier transform.

3.3 Identification of periodic regions in microsatellite sequences

We evaluated the effectiveness of the PPS method by comparing the Fourier and the PPS spectra of complex microsatellite DNA sequences. Microsatellites are abundant and distributed all over the eukaryotic genomes with variable frequency. They are used as genetic markers in forensics, parentage assessment, positional cloning and population and evolutionary genetics. The most studied groups of tandem repeats in genomes are microsatellites (patterns of 10 bp) and minisatellites (patterns of 100 bp). We studied the the PPS method on periodicity detection on Human microsatellite repeat KLK1 AC DNA (GenBank accession number: M65145, 1072 bp). The Fourier power spectrum of the DNA sequence indicates there are two high peaks corresponding to approximate 11 and 12 periodicities (Fig. 4a), but it is difficult to identify other periodicities from the Fourier power spectrum. The PPS spectrum SNR at periodicities 2–100 clearly shows that a few periodicity peaks with large spectrum SNR values in the DNA sequence (Fig. 4b). The significant SNR values for short periodicities are: $P_7: 1.3873$, $P_8: 1.2002$, $P_{11}: 1.5389$, $P_{12}: 2.4196$. Our PPS method can thus identify 4 periodicities (repeats): 7, 8, 11, and 12, which can be located using sliding window PPS (Fig. 4c). The result from sliding windows shows the different locations for these periodic signals on the DNA sequence. For example, the 11-mer repeats are located between 92 and 781 bp. Comparing the sliding window PPS SNR plot of period 11 with that of period 12 (Fig. 4b, c), we can see that the peak regions from period 11 plot and the period 12 plot are not the same. The region centered at 200 bp position only shows a periodicity of 11, but not of 12. This result suggests the period 11 is not exactly derived from a periodicity of 12. The locations the period 11 agree with previous studies (Sharma et al. 2004; Gupta et al. 2007).

The periodicities of microsatellite repeat KLK1 AC DNA were studied previously by Fourier transform (Sharma et al. 2004) and the EPSD (exactly periodic subspace decomposition) method (Gupta et al. 2007). These two repeat finding methods could only detect 2 or 11 mer repeats in the sequence, however, our PPS method can capture five periodicities 2, 7, 8, 11, and 12. To validate the existence of the additional three periodicities 7, 8 and 12 in the sequence, we examined the regions identified by sliding window PPS for corresponding periodicities (Fig. 4c). The periodic sequence patterns in these regions are illustrated in Fig. 5. The results show that the imperfect 7 mer repeats are located between 450-550 bp of the sequence, with 7 mer periodic
Fig. 4 Power spectra of DNA sequence Human microsatellite repeat KLK1 AC DNA. a Fourier power spectrum. b PPS spectrum SNR. c PPS spectrum SNRs of sliding windowed sequence for different periodicities, window size = 60 bp
Fig. 5  Periodic sequences in Human microsatellite repeat KLK1 AC DNA by the PPS.  
\[ \text{a 7-periodic sequence.} \]
\[ \text{b 8-periodic sequence.} \]
\[ \text{c 12-periodic sequence} \]

The imperfect 8 mer repeats are located between 200–300 bp of the sequence, with 8 mer periodic pattern GGGGTGGA. The imperfect 12 mer repeats are located between 50–200 bp of the sequence, with 12 mer periodic pattern TAAATTGGAGGA. These results show that the PPS method can clearly identify some periodicities that are missed in Fourier transform and the EPSD method.

The second example is periodicity detection in Human microsatellite repeats (GenBank locus:HSVDJSAT, 1985 bp) using the PPS method. This DNA sequence contains variable length tandem repeats (VLTRs) (Hauth and Joseph 2002). Figure 6a is the Fourier power spectrum of the DNA sequence and indicates the complexity and high noise level in Fourier power spectrum. Thus it is difficult to detect period-
Fig. 6  Power spectra of DNA sequence HSVDJSAT. a Fourier power spectrum. b PPS spectrum SNR. c PPS spectrum SNRs of sliding-windowed sequence for different periodicities. The window size is 100 bp.
icities from Fourier power spectrum. From the PPS spectrum in Fig. 6b, we can detect accurately the short periodicities: 4, 6, 8, 10, 22, 49, and 50. The corresponding SNR values of these periodicities are: \( P_4:1.2781, P_6:1.4620, P_8:1.0349, P_{10}:1.7631, P_{22}:1.0781, P_{49}:1.1268, P_{50}:1.1023 \). The strongest periodicity in this DNA sequence is periodicity 10, which is in agreement with previous studies (Hauth and Joseph 2002; Gupta et al. 2007). From the sliding window PPS spectra of periodicities 4, 6 and 10 in Fig. 6c, we can locate the positions of these periodicities. For example, a major location of the periodicity 4 is between 700–900 bp; the periodicity 6 is mainly located around 400 and 1000 bp, periodicity 10 is mainly located at 1180 and 1400 bp. The previous studies (Hauth and Joseph 2002; Gupta et al. 2007) can identify similar periodicities but those methods need different threshold settings and sometimes get conflicting results.

These two case studies of microsatellite repeats demonstrate that the PPS method can effectively detect many hidden periodicities and locate periodic regions and patterns of DNA sequences using the sliding window approach. It can detect a broad range of latent and complex periodicities in DNA sequences, which could be missed in other periodicity finding methods.

### 3.4 Identification of periodic regions in genomes

We tested the PPS method on capturing typical long latent periodicities in genomes. The first example is periodicity analysis of a region on chromosome I (1661021 – 1663249) of *C. elegans* (GenBank:NC_003279). This region contains a periodicity of 10 by the spectral-statistical approach (Chaley et al. 2014). The 10 bp periodicity in genome may associate with DNA structure and nucleosome positioning (Segal et al. 2006). The Fourier spectrum of the region shows multiple high peaks (Fig. 7a). It is difficult to exactly locate periodicities from the Fourier spectrum. However, the PPS analysis can clearly show the periodicities 2, 5, and 10, and locate the corresponding periodic regions by the sliding window approach (Fig. 7b, c). The results from the PPS method agree with the spectral-statistical approach (Chaley et al. 2014). However, the spectral-statistical approach could not detect periodicity 2 and 5 in this region. Without exception, this result shows that the PPS method can capture more periodicities than other periodicity finding methods, including the spectral-statistical approach.

Another genomic periodicity analysis was performed on the region (3239317 – 3239812) of chromosome IV of *A. thaliana* (GenBank:CP002687). This DNA region shows periodicity 7 from previous study (Chaley et al. 2014). The PPS method can detect the periodicity 7, and a weak periodicity 23 in this region (Fig. 8b, c), but Fourier spectrum has difficulty to accurately identify this weak periodicity 23. In addition, this weak periodicity 23 was not captured using the spectral-statistical approach (Chaley et al. 2014).

The last example is the PPS spectrum analysis of the whole chromosome I of *P. falciparum* (GenBank:NC004325, 643292 bp). The PPS spectrum shows pronounced periodicities 2, 3, 7 and 21 (Fig. 9b, c), but the Fourier spectrum cannot accurately identify these periodicities because of high background noise (Fig. 9a). For example, the periodicity 21 was found previously using Fourier transform (Nunes...
Fig. 7  Power spectra of DNA sequence on chromosome I (166102 – 1663249) of *C. elegans*. a Fourier power spectrum. b PPS spectrum SNR. c PPS spectrum SNRs of sliding-windowed sequence for the periodicities 2, 5 and 10. The window size is 60 bp.
Fig. 8 Power spectra of DNA sequence on chromosome IV (3239317 – 3239812) of *A. thaliana*. a Fourier power spectrum. b PPS spectrum SNR. c PPS spectrum SNRs of sliding-windowed sequence for the periodicities 7 and 23. The window size is 60 bp.
Fig. 9  Power spectra of the whole chromosome I of *P. falciparum*. a Fourier power spectrum. b PPS spectrum SNR. c PPS spectrum SNRs of non-overlapping sliding-windowed sequence for periodicities 7 and 21. The window size is 1000 bp.
et al. 2011), but the periodicity 7 can only be identified by the PPS method. These examples confirm that the PPS method can detect more periodicities than the Fourier transform.

3.5 Computational complexity

One of the challenging problems when applying Fourier transform to biological sequences is its high computational complexity. The fast Fourier transform (FFT) still needs $O(N \log N)$ computational time. The computation complexity of PPS for specific periodicity $P$ of protein sequence of length $N$ only involves small number of multiplications of the spectrum matrix with computational complexity $O(P^2 + P)$ and congruent derivative vector with computational complexity $O(N)$. Because periodicity $P$ is usually much smaller than $N$, $O(P^2 + P)$ is very small and can be ignored, then the computation complexity of PPS is linear with signal length, $O(N)$. Because we only need to compute power spectrum at specific periodicities in many applications, instead of full power spectrum in Fourier transform, the computational performance of the PPS power spectrum for specific periodicities is more efficient than Fourier transform. In addition, since the spectrum transform matrices consist of a constant number of entries, they can be precomputed and used directly to reduce the computational time of the PPS algorithm. Benchmarking tests were performed on an Intel Core i5-M560 processor, 2.67GHz, 6G RAM. The running time of the PPS method and MATLAB built-in FFT function for DNA sequences of different lengths is shown in Fig. 10. Considering that the FFT is a pre-compiled executable in MATLAB, while our PPS is not, the result from Fig. 10 shows that the PPS method is much faster than the FFT method. In addition, benchmark performance tests show high oscillation of execution time in FFT for long lengths. The PPS method has linear runtime for any DNA lengths as shown in Fig. 10. These results validate the complexity analysis, and demonstrate the efficiency of the PPS method.

![Fig. 10](image-url) The runtime of the PPS method and the MATLAB FFT method for DNA sequences of different lengths

\[ (a) \]
4 Discussion

Periodic regions in genomes often play important roles in genomic functions. Thor-oughly capturing periodic patterns and understanding the source of periodicity in genomic sequences thus both represent important problems in biology. The major advantages of the PPS spectrum is that it can capture all the hidden periodicities because the PPS spectrum reduces high noise level and spectral leakage as in Fourier spectrum. Using the sliding window PPS approach, we can clearly locate the regions of repeats that contribute the periodicities. The examples in this study demonstrate that the PPS method is robust in identifying and locating periodicities in coding sequences.

An open question is that a PPS spectrum becomes smooth and lacks any periodicity peaks after a certain position. It is currently unclear what is the maximum value of a periodicity that the PPS method can detect. It seems that this maximum value may depend both on periodic features as well as the length of the genomic sequences. We examine the boundary periodic length for different protein sequences, we find the position approximately equals to $\sqrt{2N}$. We will investigate this problem in our future studies.

5 Conclusions

In this paper, we propose a novel method to compute periodic power spectrum based on the periodic distribution of nucleotides of DNA sequences. The PPS method can directly identify and locate imperfect periodic patterns in DNA sequences. PPS is a deterministic method, whose strength is to detect periodicities while minimizing spectral leakage and reducing background noise in the spectrum. The PPS algorithm has been used to identify periodicities in different genomic elements, including exons, microsatellite DNA sequences, and entire chromosomes. Illustrative examples demonstrate the effectiveness of our algorithm to identify periodicities. There may be applications of PPS in the analysis of genomes, and the identification of genomic signatures.

Acknowledgments We are grateful to Professor Donald Morrison for helpful discussion, and to the anonymous reviewers for their valuable comments. We thank Emily Yin from Fremd high school for proof reading of the paper.

References

Afreixo V, Ferreira PJ, Santos D (2004) Fourier analysis of symbolic data: a brief review. Digital Signal Process 14(6):523–530
Agrawal R, Faloutsos C, Swami A (1993) Efficient similarity search in sequence databases. Springer, New York
Anastassiou D (2000) Frequency-domain analysis of biomolecular sequences. Bioinformatics 16(12):1073–1081
Arora R, Sethares WA (2007) Detection of periodicities in gene sequences: a maximum likelihood approach. In: IEEE international workshop on genomic signal processing and statistics, GENSIPS 2007. IEEE, pp 1–4
Buchner M, Janjarasjitt S (2003) Detection and visualization of tandem repeats in DNA sequences. IEEE Trans Signal Process 51(9):2280–2287

Chaley M, Kutyrkin V, Tulbashева G, Tepluhchina E, Nazipova N (2014) Heterogeneity: database of genome periodicity. Database 2014:bau040

Chechetkin V, Turygin AY (1995) Search of hidden periodicities in DNA sequences. J Theor Biol 175(4):477–494

Chen K, Meng Q, Ma L, Liu Q, Tang P, Chiu C, Hu S, Yu J (2008) A novel DNA sequence periodicity decodes nucleosome positioning. Nucleic Acids Res 36(19):6228–6236

Costa A, Melucci M (2010) An information retrieval model based on discrete Fourier transform. In: Advances in multidisciplinary retrieval. Springer, New York, pp 84–99

Datta S, Asif A (2005) A fast DFT based gene prediction algorithm for identification of protein coding regions. In: ICASSP (5), pp 653–656

Epps J, Ying H, Huttley GA (2011) Statistical methods for detecting periodic fragments in DNA sequence data. Biol Direct 6(21):1–16

Foster G (1995) The cleanest Fourier spectrum. Astron J 109:1889–1902

Glunčič M, Paar V (2013) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Res 41(1):e17–e17

Grover A, Ashwarya V, Sharma P (2012) Searching microsatellites in DNA sequences: approaches used and tools developed. Physiol Mol Biol Plants 18(1):11–19

Gupta R, Sarthi D, Mittal A, Singh K (2007) A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequences. EURASIP J Bioinform Syst Biol 2007:3–3

Gymrek M, Golan D, Rosset S, Erlich Y (2012) lobSTR: a short tandem repeat profiler for personal genomes. Genome Res 22(6):1154–1162

Hauth AM, Joseph DA (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinformatics 18(suppl 1):S31–S37

Herzel H, Weiss O, Trifonov EN (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15(3):187–193

Hoang T, Yin C, Zheng H, Yu C, He RL, Yau SST (2015) A new method to cluster DNA sequences using Fourier power spectrum. J Theor Biol 372:135–145

Illingworth CJ, Parkes KE, Snell CR, Mullineaux PM, Reynolds CA (2008) Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to gcr2, a candidate plant g PCR? Biophys Chem 133(1):28–35

Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vacser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci 89(6):2195–2199

Koning de AJ, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7(12):e1002,384

Korotkov EV, Korotkova MA, Kudryashov NA (2003) Information decomposition method to analyze symbolic sequences. Phys Lett A 312(3):198–210

Kotlar D, Lavner Y (2003) Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. Genome Res 13(8):1930–1937

Lyon DA (2009) The discrete Fourier transform, part 4: spectral leakage. J Object Technol 8(7)

Messaoudi I, Elloumi-Oueslati A, Lachiri Z (2014) Building specific signals from frequency chaos game and revealing periodicities using a smoothed Fourier analysis. IEEE/ACM Trans Comput Biol Bioinform 11(5):863–877

Murray KB, Gorse D, Thornton JM (2002) Wavelet transforms for the characterization and detection of repeating motifs. J Mol Biol 316(2):341–363

Narzisi G, Schatz M (2015) The challenge of small-scale repeats for indel discovery. Front Bioeng Biotechnol 3(8)

Nunes MC, Wanner EF, Weber G (2011) Origin of multiple periodicities in the Fourier power spectra of the plasmidium falciparum genome. BMC Genomics 12(Suppl 4):S4

Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, Van Swieten JC, Myllykangas L et al (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72(2):257–268

Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins Struct Funct Bioinform 39(2):178–194
Scargle JD (1982) Studies in astronomical time series analysis. II-statistical aspects of spectral analysis of unevenly spaced data. Astrophys J 263:835–853
Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JFZ, Widom J (2006) A genomic code for nucleosome positioning. Nature 442(7104):772–778
Shapiro JA, von Sternberg R (2005) Why repetitive DNA is essential to genome function. Biol Rev 80(02):227–250
Sharma D, Issac B, Raghava G, Ramaswamy R (2004) Spectral repeat finder (SRF): identification of repetitive sequences using fourier transformation. Bioinformatics 20(9):1405–1412
Silverman B, Linsker R (1986) A measure of DNA periodicity. J Theor Biol 118(3):295–300
Sutherland GR, Richards RI (1995) Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci 92(9):3636–3641
Suvorova YM, Korotkova MA, Korotkov EV (2014) Comparative analysis of periodicity search methods in DNA sequences. Compt Biol Chem 53:43–48
Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R (1997) Prediction of probable genes by fourier analysis of genomic sequences. Bioinformatics 13(3):263–270
Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13(1):36–46
Trifonov EN (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A Stat Mech Appl 249(1):511–516
Voss R (1992) Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. Phys Rev Lett 68:3805–3808
Wang J, Liu G, Zhao J (2012) Some features of Fourier spectrum for symbolic sequences. Numer Math A J Chin Univ 4:341–356
Wang L, Stein LD (2010) Localizing triplet periodicity in DNA and cDNA sequences. BMC Bioinform 11(1):550
Wang W, Johnson DH (2002) Computing linear transforms of symbolic signals. IEEE Trans Signal Process 50(3):628–634
Welch PD (1967) The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 15(2):70–73
Wojcik EA, Brzostek A, Bacolla A, Mackiewicz P, Vasquez KM, Korycka-Machala M, Jaworski A, Dziadek J (2012) Direct and inverted repeats elicit genetic instability by both exploiting and eluding DNA double-strand break repair systems in mycobacteria. PloS One 7(12):e51–e64
Yin C (2015) Representation of DNA sequences in genetic codon context with applications in exon and intron prediction. J Bioinform Comput Biol 13(2):1550004
Yin C, Yau SST (2005) A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. J Comput Biol 12(9):1153–1165
Yin C, Yau SST (2007) Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol 247(4):687–694
Yin C, Yau SST (2015) An improved model for whole genome phylogenetic analysis by Fourier transform. J Theor Biol 359(21):18–28
Yin C, Chen Y, Yau SST (2014a) A measure of DNA sequence similarity by fourier transform with applications on hierarchical clustering. J Theor Biol 359:18–28
Yin C, Yin XE, Wang J (2014b) A novel method for comparative analysis of DNA sequences by Ramanujan–Fourier transform. J Comput Biol 21(12):867–879