Journal of Theoretical Biology 412 (2017) 138–145

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Identification of repeats in DNA sequences using nucleotide distribution uniformity

Changchuan Yin

Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607-7045, USA

ARTICLE INFO

Keywords:
Repetitive elements
DNA sequence
Genome
Distribution uniformity

ABSTRACT

Repetitive elements are important in genomic structures, functions and regulations, yet effective methods in precisely identifying repetitive elements in DNA sequences are not fully accessible, and the relationship between repetitive elements and periodicities of genomes is not clearly understood. We present an *ab initio* method to quantitatively detect repetitive elements and infer the consensus repeat pattern in repetitive elements. The method uses the measure of the distribution uniformity of nucleotides at periodic positions in DNA sequences or genomes. It can identify periodicities, consensus repeat patterns, copy numbers and perfect levels of repetitive elements. The results of using the method on different DNA sequences and genomes demonstrate efficacy and accuracy in identifying repeat patterns and periodicities. The complexity of the method is linear with respect to the lengths of the analyzed sequences. The Python programs in this study are freely available to the public upon request or at https://github.com/cyinbox/DNADU.

1. Introduction

Repetitive elements in DNA sequences consist two or more copies of approximate patterns of nucleotides and are abundant in both prokaryotic and eukaryotic genomes. Over two-thirds of the human genome and 5–10% bacterial genomes are repetitive regions (de Koning et al., 2011). Repetitive elements play important roles in genome structure and functions such as nucleoprotein complex formation, chromosome structure, and gene expression. Various diseases including cancer and neurodegenerative disease can also arise from changes of repetitive elements. The distribution of repetitive DNA sequences can be used as fingerprints of bacterial genomes (Versalovic et al., 1991) and human individuals.

Repetitive elements are complex structures. They may exist as imperfect tandem repeats, insertion and deletions in repeats, interspersed repeats, palindromic sequences, etc. These partial and hidden repeat signals in DNA sequences are difficult to analyze through straightforward observation and sequence comparison.

Currently, repetitive elements and hidden periodicities of DNA and protein sequences are primarily detected by digital signal processing and statistical approaches (Treangen and Salzberg, 2011). In most signal processing methods, DNA sequences are converted to numerical sequences, and the hidden periodicities arising from repetitive elements can be identified by Fourier power spectrum at specific periodicities (Yin and Wang, 2016). Commonly used signal processing methods by Fourier transform include SRF maps (Sharma et al., 2004), spectral analysis (Buchner and Janjarasjitt, 2003), Ramanujan–Fourier transform (Yin et al., 2015), and the periodic power spectrum method (Yin and Wang, 2016). The statistical methods are based on distribution analysis of nucleotides in DNA sequences. The common statistical methods for repeat findings are tandem repeats finder (Benson, 1999) and statistical spectrum (Epps et al., 2011), maximum likelihood estimation (Arora and Sethares, 2007), the Monto Carlo method (Chaley et al., 1999; Korotkova et al., 1999), and information decomposition (Korotkov et al., 2003). Besides signal processing and statistical approaches, sequence alignments such as RepeatMask are also used to identify repetitive patterns in genomes (Smit et al., 1996), and require known reference repeat sequences.

Despite significant advances in repeat finding, it is still difficult to precisely capture the essential features of repetitive elements such as consensus patterns, perfect levels and copy numbers of repeats. For example, while Fourier transform is the most common used approach for finding repeats, it may not exactly correlate the strength of Fourier power spectrum with the perfect level of repeat patterns. Furthermore, since Fourier power spectrum is weak for short DNA sequences and long harmonious periodicities are embedded in short periodicities, Fourier transform cannot capture repeats in short DNA sequences and long harmonious periodicities. Moreover, the relationship between repetitive elements and periodicities of genomes is not fully understood. The statistics based methods often require different parameter settings and known reference repeat databases. Thus there is a high potential for improving the accuracy for identifying repetitive elements

E-mail address: cyin1@uic.edu.

http://dx.doi.org/10.1016/j.jtbi.2016.10.013
Received 14 August 2016; Received in revised form 10 October 2016; Accepted 27 October 2016
Available online 02 November 2016

and better understanding the relationship of periodicities and repeats in DNA sequences (Suvorova et al., 2014; Epps et al., 2011; Illingworth et al., 2008).

In this paper, we present an *ab initio* method to quantitatively identify repetitive sequences and periodicities in DNA sequences. The method is based on the nucleotide distribution uniformity at periodic positions in DNA sequences or genomes. The distribution uniformity of nucleotides reflects the unbalance of nucleotide frequencies on periodic positions and thus can indicate the strength for periodic signals in DNA sequences. The method can also reveal the consensus repeat pattern for the major periodicity of DNA sequences, and quantitatively determine the perfect level and copy numbers of repetitive sequences. The proposed method also formulates the relationship between repetitive elements and the corresponding periodicities in DNA sequences.

2. Methods and algorithms

2.1. Periodic nucleotide frequencies in a DNA sequence

A DNA molecule consists of four linearly joined nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G). A DNA sequence can be represented as a string of the four characters A, T, C, and G. The nucleotide frequencies at periodic positions, which can be represented by a congruence derivative vector (Wang et al., 2012; Yin and Wang, 2016; Wang and Yin, 2016), reflect the arrangement of repetitive elements and inner periodicities in a DNA sequence. The congruent derivative vector of a nucleotide α for a specific periodicity is constructed by the cumulative occurring frequencies of nucleotides at the periodic positions (Definition 2.1).

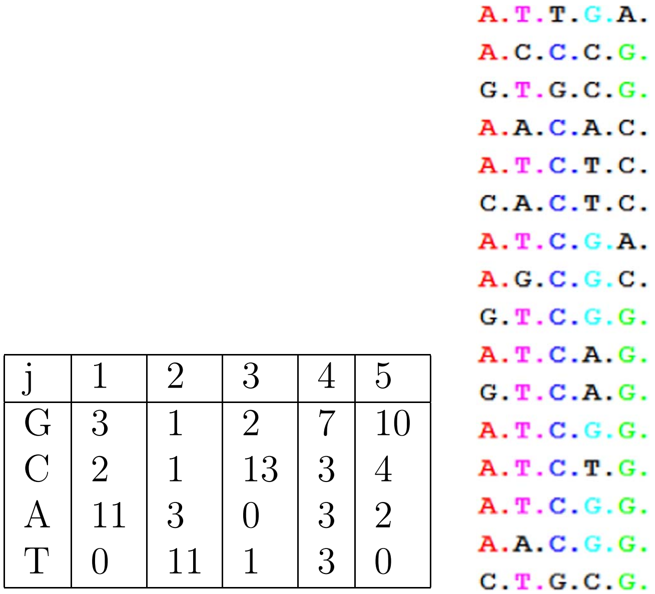
Definition 2.1. For a DNA sequence of length n, let $u_\alpha(k) = 1$ when the nucleotide α appears at position k, otherwise, $u_\alpha(k) = 0$, where $\alpha \in \{A, T, C, G\}$ and $k = 1,\ldots,n$. The congruence derivative vector of the nucleotide α of the sequence for periodicity p, is defined as

$$f_{\alpha,j} = \sum_{\mathrm{mod}(k,p)=j} u_\alpha(k)j = 1,\ldots,p, \quad k = 1,\ldots,n \tag{1}$$

where $\mathrm{mod}(k, p)$ is the modulo operation and returns the remainder after division of k by p, and $f_\alpha = (f_\alpha(1), f_\alpha(2),\ldots,f_\alpha(p))$.

Four congruence derivative vectors f_α of periodicity p for nucleotides A, T, C and G form a congruence derivative (CD) matrix of size $4 \times p$. The columns of the CD matrix indicate nucleotide frequencies at the periodic positions $k = pt - q$, where k is the position index of a DNA sequence, $t = 1, 2, \ldots,$ and $q = p - 1,\ldots,2, 1, 0$. For example, consider the CD matrix of periodicity 5 for DNA sequence, the first column of the CD matrix shows the nucleotide frequencies at periodic positions $k = 1, 6, 11,\ldots,5t - 4$; the second column of the matrix shows the nucleotide frequencies at periodic positions $k = 2, 7, 12,\ldots,5t - 3$; the third column of the matrix shows the nucleotide distributions at periodic positions $k = 3, 8, 13,\ldots,5t - 2$, and so on. In this way, the CD matrix of a DNA sequence describes nucleotide frequencies at all periodic positions and can be used to efficiently compute Fourier power spectrum and determine periodicities in the DNA sequence (Yin and Wang, 2016). In this study, instead of Fourier transform, we use the CD matrix to identify repetitive elements and periodicities directly. This approach offers elaboration of the repetitive elements such as the consensus repeat pattern, copy number and perfect level.

To illustrate the nucleotide frequencies in the CD matrix,an artificial DNA sequence of 80 bp, ATTGAACCCGGTGCGAACACATCTCCACTCAT CGAAGCGCGTCGGATCAGGTCAGATCGGATCTGATCGGAACGGCTGCG, which contains 5 bp approximate repeats, is constructed. The CD matrix of periodicity 5 for this sequence is shown in Fig. 1. The base with the highest frequency in each column is labeled in color, while the other bases are labeled in black. The consensus sequence of the repeats can be identified as ATCGG from the highest frequency in each column in the CD matrix. Thus, we may determine that the DNA sequence contains 16 copies of approximate ATCGG (Fig. 1). The bases shown in black in columns are then the mismatched nucleotides when the actual sequence is compared with the consensus sequence at the periodic positions.



j	1	2	3	4	5
G	3	1	2	7	10
C	2	1	13	3	4
A	11	3	0	3	2
T	0	11	1	3	0

A.T.T.G.A.
A.C.C.C.G.
G.T.G.C.G.
A.A.C.A.C.
A.T.C.T.C.
C.A.C.T.C.
A.T.C.G.A.
A.G.C.G.C.
G.T.C.G.G.
A.T.C.A.G.
G.T.C.A.G.
A.T.C.G.G.
A.T.C.T.G.
A.T.C.G.G.
A.A.C.G.G.
C.T.G.C.G.

Fig. 1. The congruence derivative matrix of periodicity 5 of an artificial DNA sequence (left) and the inferred 16 copies of approximate 5 bp repeats (right). The consensus sequence of the repeats inferred from the matrix is ATCGG in color. The DNA sequence of 80 bp is described in the text. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

2.2. Computing the distribution uniformity (DU) of a DNA sequence

Identification of repeats and periodicities in DNA sequences is often realized by Fourier transform. However, Fourier power spectrum may not characterize repetitive elements precisely. It does not identify the consensus repetitive pattern, copy number and perfect level. We propose to employ the periodic nucleotide frequencies as a measure of repeats and periodicities. Previously we showed that the nucleotide distribution on periodic positions correlates with the strength of periodicity 3 (Yin and Yau, 2005). We here generalize to use the nucleotide distribution as the signal measure for any periodicities. Since the CD matrix contains the nucleotide frequencies on periodic positions, the variance of the matrix elements can measure the nucleotide distribution. For the CD matrix of periodicity p, the summation of total $4p$ elements of the matrix is equal to length n of the DNA sequence, the mean of the elements of the matrix is $\frac{n}{4p}$. To measure the nucleotide distribution and use it to indicate the periodic signal in a DNA sequence, we define the distribution uniformity (DU) of a DNA sequence using the CD matrix (Definition 2.2).

Definition 2.2. For a DNA sequence of length n, let $f_{i,j}$ be an element of the CD matrix of periodicity p, the distribution uniformity of periodicity p of the sequence is defined as

$$DU(p) = \sum_{i=1}^{4}\sum_{j=1}^{p}\left(f_{i,j} - \frac{n}{4p}\right)^2 \tag{2}$$

From Definition 2.2., we notice that the distribution uniformity at periodicity p is an intuitive description for the level of unbalance of nucleotide frequencies on periodic positions. It depends on the quadratic function of the nucleotide frequencies, sequence and periodicity length. Additionally, Definition 2.2 can be rewritten as Eq. (3). It shows that for fixed length and periodicity, the distribution uniformity only depends on the squares of nucleotide frequencies. The proof of Eq. (3) is provided in the supplementary materials:

$$DU(p) = \sum_{i=1}^{4}\sum_{j=1}^{p} f_{i,j}^2 - \frac{n^2}{4p} \tag{3}$$

For a DNA sequence of length n consisting of tandem perfect repeats of size p, each column of the CD matrix contains three zeros and one non-zero element n/p. Therefore, the distribution uniformity of tandem perfect repeats is described in the following equation:

$$DU_X(p) = \sum_{i=1}^{4}\sum_{j=1}^{p}\left(f_{i,j} - \frac{n}{4p}\right)^2 = p\left(\left(\frac{n}{p} - \frac{n}{4p}\right)^2 + 3\left(0 - \frac{n}{4p}\right)^2\right) = \frac{3n^2}{4p} \tag{4}$$

Definition 2.2 and Eq. (4) indicate that if the length n of a DNA sequence is long or if repeat length p is short, then the nucleotide distribution uniformity becomes large. For a random DNA sequence with uniform nucleotide distribution, each element of the CD matrix of periodicity p is equal to $\frac{n}{4p}$, so the $DU(p)$ is zero. Thus the distribution uniformities of any DNA sequences are in range $\left[0, \frac{3n^2}{4p}\right]$.

A copy number of repeats indicates how many repeats are there in a DNA sequence or genome. Copy number variations depend on repeat types and genomes and play an important role in generating variation in population and disease phenotype (McCarroll and Altshuler, 2007). The strength of a distribution uniformity is impacted by the copy number of repeats. The more copies of repeats, the stronger of the distribution uniformity. We define the copy number of repeats as follows (Definition 2.3).

Definition 2.3. The copy number of the repeats of size p is equal to the division of the sequence length by the periodicity length p:

$$CP(p) = \frac{n}{p} \tag{5}$$

The above definitions indicate that the longer a DNA sequence is, the larger the copy number is and the stronger distribution uniformity is. To consistently compare the strengths of periodic signals in DNA sequences of different lengths, we use the normalized distribution uniformity (NDU), which can be considered as the mean distribution uniformity. It is defined as the distribution uniformity of periodicity p divided by the length of the sequence (Definition 2.4). $NDU(p)$ can be used to indicate the existence of the periodicity p in a DNA sequence.

Definition 2.4. The normalized distribution uniformity (NDU) is equal to the distribution uniformity divided by the sequence length n:

$$NDU(p) = \frac{DU(p)}{n} \tag{6}$$

From Eq. (3), the normalized distribution uniformity of periodicity p of a DNA of length n can be further described as in the following equation:

$$NDU(p) = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{p} f_{i,j}^2 - \frac{n}{4p} \tag{7}$$

For tandem perfect repeats, we get NDU by Definitions 2.3 and Eq. (4) as follows:

$$NDU_X(p) = \frac{DU_X(p)}{n} = \frac{3n^2}{4pn} = \frac{3n}{4p} = \frac{3}{4}CP(p) \tag{8}$$

Eq. (8) indicates that if the NDU of tandem perfect repeats is larger than 1, the copy number of tandem perfect repeats is at least 4/3. This fact designates the sequence feature of the NDU threshold value as 1. When the NDU of a DNA sequence is larger than 1, it suggests existence of a significant periodicity and repeats in the sequence. This is the foundation of our algorithm for detecting periodicity and repeats by using the distribution uniformity of nucleotides.

2.3. Identifying the repeat pattern of length p *in a DNA sequence*

Using the CD matrix of periodicity p, we can determine the dominant nucleotide in each column of the matrix, where the nucleotide has the maximum frequency at the qth periodic position $q = 1, 2\ldots, p$ (Definition 2.5). From the positions of dominant nucleotides, we can determine the consensus repeat pattern.

Definition 2.5. The dominant nucleotide of the congruence derivative matrix of periodicity p of a DNA sequence is the element with the maximum frequency in each column of the matrix.

$$d_j = \max(f_{i,j}), \quad i \in \{1, 2, 3, 4\}, \quad j = 1, 2, \ldots, p \tag{9}$$

The consensus repeat pattern of size p of a DNA sequence may be inferred from the dominant nucleotide frequencies of the congruence derivative matrix of a periodicity p. In a CD matrix, d_j is the maximum of the jth column, if the corresponding nucleotide in the matrix for d_j is $\alpha \in \{G, C, A, T\}$, then nucleotide α appears periodically at the jth positions in the sequence. In detail, let r_j be jth nucleotide of the repeat unit of size p of the DNA sequence, $j = 1, 2, \ldots p$. If $d_j = f_{i,j}$, then $r_j = \alpha_i$, $\alpha \in \{G, C, A, T\}$. Thus the consensus repeat pattern can be identified.

After the consensus repeat pattern of repetitive elements is determined, a parameter is needed to measure the perfect level of the approximate repeats compared with the tandem perfect repeats of the same length. We define the perfect level $PR(p)$ as the percentage of matched nucleotides in a DNA sequence with tandem perfect repeats of the same length (Definition 2.6).

Definition 2.6. The perfect level $PR(p)$ is the percentage of matched nucleotides in a DNA sequence with tandem perfect repeats of the same length. It is equal to the summation of dominant column elements in the CD matrix divided by the sequence length n:

$$PR(p) = \frac{1}{n}\sum_{j=1}^{p} d_j \tag{10}$$

Since the NDU value of approximate repeats also depends on perfect level, if at least two copies of tandem repeats are required, then from Eqs. (3)–(10), the perfect level must be at least 2/3 (i.e., 66.67%) to make NDU threshold larger than 1. In this study, we use a NDU threshold of 1 (i.e., 100%) to determine whether a distribution uniformity truly indicates the periodicity.

As an example, from the CD matrix of periodicity 5 in Fig. 1, the dominant nucleotide frequencies d_j of the DNA sequence of length 80 bp are [11,11,13,7,10] and the derived consensus repeat pattern is ATCGG. The perfect level computed by Eq. (10) is 65%. It indicates that among repeat elements, 65% nucleotides match 16 copies of tandem perfect repeats ATCGG.

2.4. Algorithms

The inputs for computing the distribution uniformity, copy number, and perfect level are the sequence length n and periodicity length p. Given a DNA sequence and periodicity p as inputs, we have the following algorithm, named the distribution uniformity (DU) method, for identifying repetitive elements and periodicities in a DNA sequence.

Algorithm 1. The algorithm for identifying repeats and periodicity in a DNA sequence.

```
Input: A DNA sequence of length $n$, periodicity $p$
Output: DU, NDU, consensus repeat pattern, perfect level, copy
  number
Step: initialization
1. Convert DNA sequence into 4D binary indicators.
2. Compute congruence derivative (CD) matrix of periodicity $p$ from
  the 4D indicators.
3. From the CD matrix, we can compute:
3.1. $DU(p)$ at periodicity $p$ (Eq. (3)).
3.2. $NDU(p)$ (Eq. (6)).
3.3. Consensus repeat pattern of size $p$ (Eq. (9)).
3.4. Perfect level $PR(p)$ (Eq. (10)).
3.5. Copy number for the repeat of size $p$ (Eq. (5)).
```

To compute distribution uniformities of different periodicities of a DNA sequence, we first scan the sequence in different periodicity sizes, construct the congruence derivative matrix of each periodicity, and compute the distribution uniformities of these periodicities. The periodicity with the maximum distribution uniformity reflects the dominant pattern of repetitive elements.

Just like the short-time Fourier transform (STFT), to identify the repeat regions in a DNA sequence, we adopt a fixed size window to slide along DNA sequences, and compute the NDU values of different periodicities in the window continuously. The NDU values of periodicities indicate the perfect levels and copy numbers of corresponding repeat regions. The size of a sliding window often determines the resolution of location of periodic elements. A short window can achieve high resolution of locations, but produce weak NDU signal. Because uniform distributed DNA sequences have perfect level 25%, and typical non-uniform distributed DNA sequences have perfect level 33% from our benchmark testing, hence at least 5 copies of DNA repeats are needed to generate enough DU signal with threshold value as 1 (Eqs. (3)–(10)). Therefore, the sliding window size can be 5 times of the length of latent periodicity or repeat.

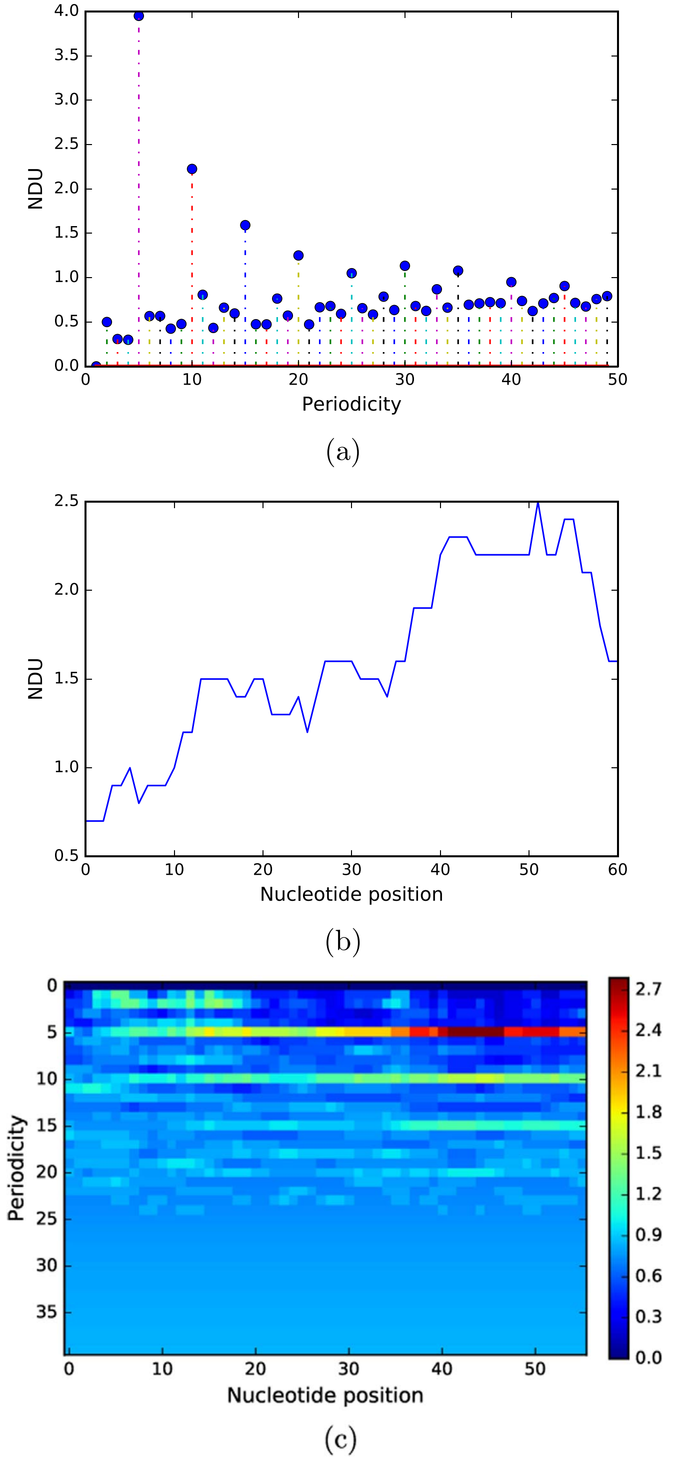
3. Results

3.1. Periodic analysis of short DNA sequences

For the artificial DNA sequence as described in Fig. 1, the DU method can identify harmonious periodicities 5, 10, 15, 20, 25, etc. (Fig. 2(a)). The NDU for periodicity 5 (Fig. 2(b)) and NDU values 1–40 in the sequence using short sliding window of length 20 bp (Fig. 2(c)) conform the repeat structures in the sequence. The results show that the signal of periodicity 5 is clear with very short window size 20 bp. The capacity of capturing periodic signals from short DNA sequences is an advantage of the proposed method.

3.2. Periodic analysis of highly complex repeats in DNA sequences

The effectiveness of the proposed algorithm is tested on *Homo sapiens* collagen type IV alpha 6 chain (COL4A6) (GenBankID:NC_001847, 6618 bp). This gene contains repeats of length 9 bp by the EPSD method (Gupta et al., 2007). The result by the DU method shows that the sequence contains repeats of 3 bp, 6 bp, 9 bp (Fig. 3(a)). The sliding window approach shows the location and strength of the periodicity 9 (Fig. 3(b)) and periodicities 1100 (Fig. 3(c)). The corresponding periodicity strength, NDU, perfect level, consensus pattern, and copy number can be revealed by the method (Table 1). Table 1 shows that the perfect levels of periodicities 3, 6 and 9 are similar, but the periodicity 3 has the largest mean NDU because of the highest copy number of periodicity 3. The periodicities in this DNA sequence identified by the proposed method are consistent with the EPSD method (Gupta et al., 2007). The results demonstrate the effectiveness of the proposed method in capturing different repeats and periodicities.



(a)

(b)

(c)

Fig. 2. The distribution uniformity of the artificial DNA sequence. The sequence and its CD matrix are shown in Fig. 1. (a) NDU of different periodicities. (b) NDU of periodicity 5 of the DNA sequence by sliding window of 20 bp. (c) NDU of different periodicities in the DNA sequence by sliding window of 20 bp.

To test if the DU method can identify complex repeats, we use the method to analyze Human microsatellite repeats (GenBank locus:HSVDJSAT, 1985 bp). The DNA sequence contains variable length tandem repeats (VLTRs) (Hauth and Joseph, 2002). The DU method may detect accurately all different the short periodicities and corresponding perfect levels and copy number of the repeats in the full DNA sequence (Fig. 4(a) and Table 2). As an example, the sliding window approach for the NDU of periodicity 7 indicates that the 7 bp

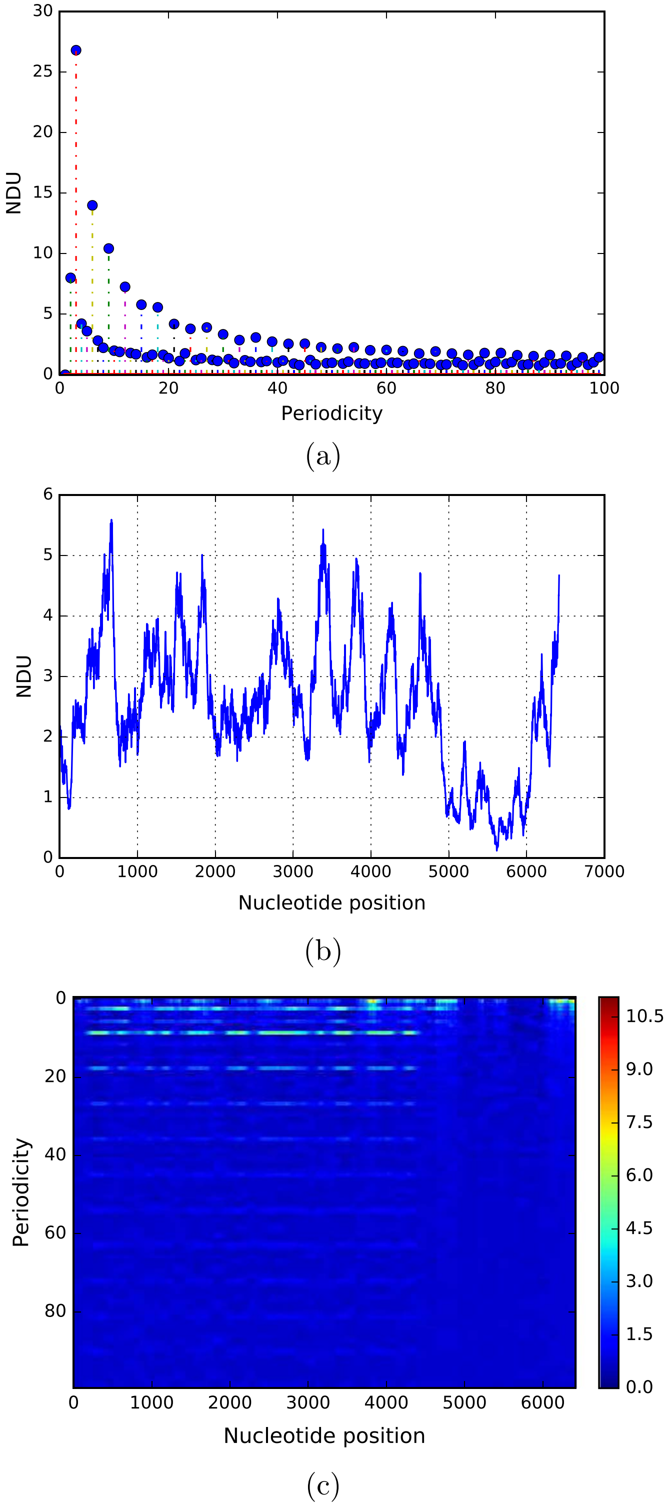


Fig. 3. The distribution uniformity analysis of *Homo sapiens* collagen type IV alpha 6 chain (COL4A6) (GenBankID:NC_001847). (a) NDU of different periodicities. (b) NDU of periodicity 9 of the DNA sequence the sliding window approach. (c) NDU of different periodicities in the DNA sequence by sliding window approach. The window size is 200 bp.

Table 1
Characterization of the repeats in human COL4A6 gene (GenBankID:NC_001847).

Periodicity	*NDU*(*p*)	*PR*(*p*)	Pattern	Copies
3	26.81036	0.3262	GGA	2206
6	13.9946	0.3262	GGAGGA	1103
9	10.4310	0.3312	GCAGGAGGT	735

repeats are between positions 1000 bp and 1550 bp when using the NDU threshold as 1 (Fig. 4(b)). The method may locate the positions of other repeats from the corresponding periodicities (Fig. 4(c)). Specially, in the region between 1000 bp and 1550 bp, two long repeats of 18 bp and 19 bp can be identified using the DU method (Table 3). These results agree with the previous studies by statistical methods (Hauth and Joseph, 2002; Gupta et al., 2007; Benson, 1999), however, the previous studies need different threshold settings and sometimes get conflicting results.

3.3. Long periods and repeats in DNA sequences

We further assess the DU method for identifying long latent periodicities and repeats in DNA sequences. These DNA sequences contain long latent periodicities and repeats, which were discovered by the Monte Carlo method (Chaley et al., 1999). The long latent periodicities and repeats identified by the DU method are listed in Table 4, with an example shown in Fig. 5. The periodicities of larger than 10 bp are assessed and the harmonious periodicities of 3 bp are not considered as long latent periodicities in the identification. The periodicities and repeats captured by the DU method in Table 4 are in agreement with the previous study (Chaley et al., 1999). Furthermore, the DU method can detect some additional long latent periodicities that were not reported in previous studies. For example, *B. subtilis* mcp A gene (GenBank: L29189) contains typical 21 bp periodicities as discovered in Chaley et al. (1999), the DU method can capture this pronounced periodicity, and also detect typical periodicities of 10 bp and 11 bp in the gene. Another example in Table 4 is the identification of multiple long periodicities in *C. thermocellum* gene for S1 protein (GenBank: X67406, i.e., S43430). The DU method can identify four long periodicities of 123 bp, 141 bp, 165 bp, and 246 bp (Fig. 5(a)). These long latent periodicities can be identified as the local maxima of the NDU values of harmonious periodicities of 3 bp (Fig. 5 (a) and (b)). The DU method shows that the periodicity of 297 bp has perfect level 0.5104 and is similar to the periodicity of 246 bp. Because the periodicity of 246 bp is shorter than the periodicity of 297 bp, the periodicity of 246 bp is a high periodic signal in the periodicity analysis. This result explains the discrepancy that this gene contains periodicity of 297 bp in previous study (Chaley et al., 1999). It is noted that the periodicity of 21 bp is popular in these gene sequences. This special periodicity of 21 bp resulted from the repetitive arrangement of 7-amino acid residues of protein, which forms the alpha-helix structure of 3.6 amino acids (Chaley et al., 1999). Another interesting observation in the results (Fig. 5(b)) is that the DU method can demonstrate the harmonious periodicities of 3 bp in these DNA sequences. The harmonious periodicities show decreasing order, and is inversely proportional to periodicity length. The long latent periodicities can then be identified from the local maxima from these harmonious periodicities. Therefore, this result demonstrates that the DU method can identify long latent periodicities in a DNA sequence and provide insightful information on the periodic signals in DNA sequences.

3.4. Computational complexity

The computation complexity of the DU method for specific periodicity p of DNA sequence of length n only involves computing CD matrix $O(n)$ and small number of multiplication of entries in the matrix $O(p^2 + p))$. Because periodicity p is much smaller than n, $O(p^2 + p))$ is very small, the computation complexity of the DU method is approximately linear with the sequence length, $O(n)$. In contrast, using Fourier transform in detecting repetitive elements has high computational complexity. The fast Fourier transform (FFT) needs $O(n \log n)$ computational time. Because only distribution uniformities at specific peri-

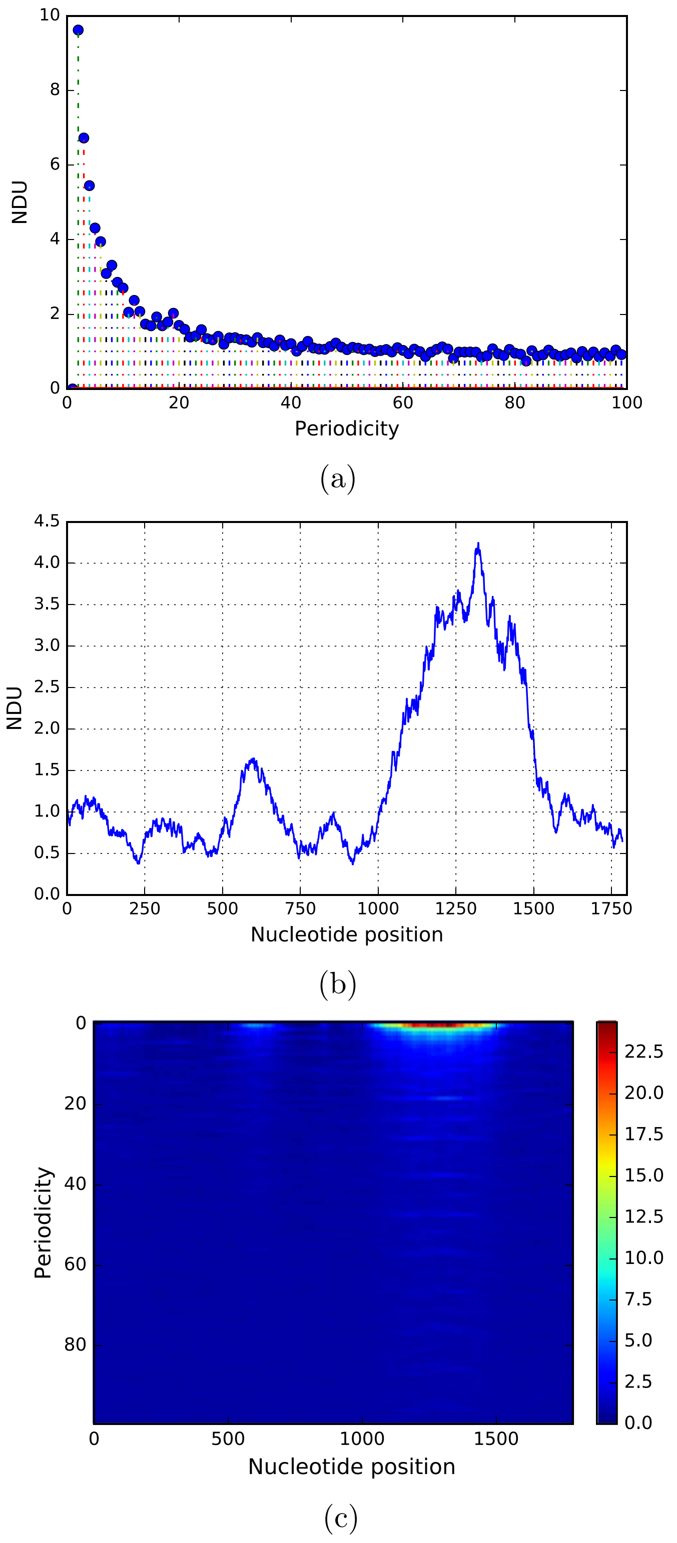


Fig. 4. The distribution uniformity analysis of variable length tandem repeats (VLTRs). (a) NDU of different periodicities. (b) NDU of periodicity 5 of the DNA sequence by the sliding window approach. (c) NDU of different periodicities of the DNA sequence by the sliding window approach. The window size is 200 bp.

odicities are needed for detecting repetitive elements of interested lengths, instead of DUs for all periodicities, the performance of the DU method for specific periodicities is more efficient than Fourier transform. Performance tests were performed on a PC with configuration as Intel Core i5 processor, 6 GB RAM. The DU method has linear runtime for any sequence lengths as shown in Fig. 6. This result verifies the complexity analysis, showing high efficiency of the DU method.

Table 2
Characterization of repeats in the human microsatellite sequence (GenBank locus: HSVDJSAT, region 1-1985 bp).

Periodicity	*NDU*(*p*)	*PR*(*p*)	Pattern	Copies
7	3.0935	0.3274	GGGGGGG	283
8	3.3153	0.3314	GGAGGGTG	248
9	2.8575	0.3312	GGGGGGGGA	220
10	2.7049	0.3314	GGAGTGGGGG	198
16	1.9299	0.3324	GGAAGGTGGGAGGGTG	124
17	1.6939	0.3324	GGAAGGTGGGAGGGTG	116
18	1.7932	0.3340	AGGGGGAGGGGGGAGGGA	110
19	2.0291	0.3420	GGCGGGGGTAGGCGGGGAG	104

Table 3
Characterization of long repeats in the human microsatellite sequence (GenBank locus: HSVDJSAT, region 1000-1550 bp).

Periodicity	*NDU*(*p*)	*PR*(*p*)	Pattern	Copies
18	1.9875	0.4560	GGGGGGGGGGGGGGGGGG	27
19	3.5210	0.5200	CTGGGAGGGCTGGGAAAGG	26

Table 4
Long periodicities and repeats identified by the DU method.

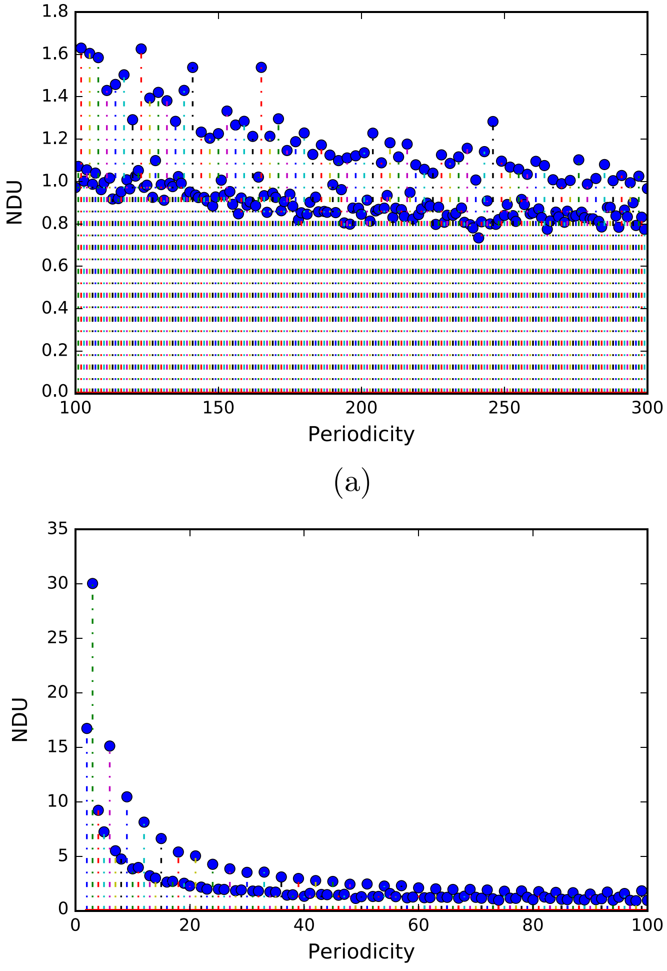
GenBank	Name	Periodicities (bp)	NDU	PR
M19713	Human tropomyosin	21, 42, 63,84	84-bp:1.6430	84-bp:0.4833
U35637	Human nebulin	21, 42, 63, 105	105-bp:6.3334	105-bp:0.3722
L29189	*B. subtilis* mcp A	10, 11, 21, 42	21-bp:6.9936	21-bp:0.3426
X67406	*C. thermocellum* S1	51, 99, 123, 141,165, 246	246-bp:1.2831	246-bp:0.5161
X67506	*C. thermocellum* anc A	129, 147, 258, 276, 294	294-bp:1.6883	294-bp:0.3952
X60999	*E. coli* rhsD	21, 42, 63, 93	93-bp:1.3807	93-bp:0.33283
M28232	*E. coli* tolA	12, 21, 33, 66	33-bp:2.8923	33-bp:0.3720

4. Discussion

Using the distribution uniformity of nucleotides in DNA sequences, we establish an *ab initio* method in identifying repeats and periodicities in the DNA sequences. The distribution uniformity intuitively depicts the level of unbalance of nucleotide frequencies on periodic positions. The unbalance of nucleotide frequencies determine the nature of repeats and periodicities in DNA sequences. The method does not rely upon previously known repeat sequence information and is natural, effective and convenient. It can quantitatively describe a DNA repeat with NDU, consensus repeat pattern, perfect level, and copy number. This method is fast with linear computational complexity and does not depend on parameter settings and statistical training as in other repeat finding methods. In addition, the method can effectively capture long latent periodicities and repeats, and retain the periodical signals for harmonious periodicities.

The proposed DU method can identify multiple repeat patterns of different lengths or the same length in a sequence. If the repeat patterns have different lengths, the method produces multiple NDU peaks at different periodicities. If these patterns have the same length, these repeat patterns contribute the strength of the same periodicity, the method can then find the locations of these repeats using sliding window approach and also get the repeat sequence patterns for detailed examination.

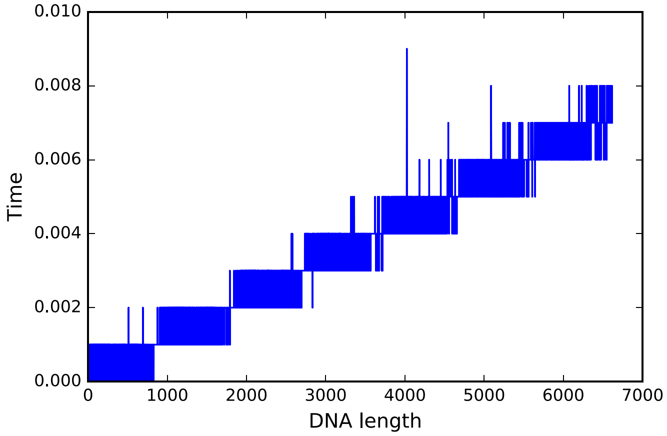
This study shows that the periodicities of genome is contributed by the tandem repeats of the corresponding length. For example, a strong periodicity of length 5 is caused by a large number of copies of 5 bp



(a)

(b)

Fig. 5. The distribution uniformity analysis of *C. thermocellum* protein S1 gene. (a) The NDU values between the periodicities of 100 and 300. (b) The NDU values between the periodicities of 2 and 100.



(a)

Fig. 6. Running time of the DU method on DNA sequences of different lengths.

repeats. The approximate repeats render high frequencies of nucleotides at periodic positions in DNA sequences. The consensus repeat patterns identified by the proposed method can be considered as the original sequences from which the approximate repeats have been evolved in evolutionary history. The perfect level of the approximate repeats represent the evolutionary progress. Thus the repetitive sequences can be used in phylogenetic analysis (Versalovic et al., 1991).

Unlike the Fourier transform method, the DU method can reveal long harmonious periodicities or capture periodicities from short DNA sequences, which are often missed in Fourier transform. The DU method also directly computes the distribution uniformity for interested periodicities, whereas Fourier transform computes the power spectrum for frequencies, and there is no a straightforward solution to convert it into the power spectrum for periodicities.

We note that DU is related to mutual information (MI) (Kullback, 1997). DU describes the unbalance level of nucleotide distributions on periodic positions, to this end, in the definition of DU, the mean value of nucleotide frequency is subtracted from the nucleotide frequencies. MI quantifies the amount of information that can be obtained from nucleotide X about another nucleotide Y that is located k positions downstream from X (Grosse et al., 2000; Korotkov et al., 2003). The relationship between DU and MI, and its application in repeat finding will be our future study.

Despite the efficacy in capturing repeats and the corresponding features by the proposed method, there are some limitations that the DU method cannot solve. One limitation is that it cannot identify interspersed repetitive elements or repeats interrupted by deletion mutations. We will address this limitation in our future research.

The frequencies of nucleotides vary significantly across eukaryotic genes and may present specific regulatory motifs (Louie et al., 2003). Because the nucleotide distribution represents the essential characteristics of a DNA sequence, the distribution uniformities of different periodicities may have broad applications in functional analysis of genomes.

Acknowledgment

We appreciate Professor Jiasong Wang at Department of Mathematics, Nanjing University, China, for his mentorship and valuable advice for this research. We are grateful to Emily Yin at Indiana University and Tung Hoang for proof reading of the paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2016.10.013.

References

Arora, R., Sethares, W.A., 2007. Detection of periodicities in gene sequences: a maximum likelihood approach. In: IEEE International Workshop on Genomic Signal Processing and Statistics, 2007. GENSIPS 2007. IEEE, New York. pp. 1–4.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27 (2), 573.

Buchner, M., Janjarasjitt, S., 2003. Detection and visualization of tandem repeats in DNA sequences. IEEE Trans. Signal Process. 51 (9), 2280–2287.

Chaley, M.B., Korotkov, E.V., Skryabin, K.G., 1999. Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples. DNA Res. 6 (3), 153–163.

de Koning, A.J., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7 (12), e1002384.

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

Grosse, I., Herzel, H., Buldyrev, S.V., Stanley, H.E., 2000. Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E 61 (5), 5624.

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.

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

Illingworth, C.J., Parkes, K.E., Snell, C.R., Mullineaux, P.M., Reynolds, C.A., 2008. Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR? Biophys. Chem. 133 (1), 28–35.

Korotkov, E.V., Korotkova, M.A., Kudryashov, N.A., 2003. Information decomposition method to analyze symbolical sequences. Phys. Lett. A 312 (3), 198–210.

Korotkova, M.A., Korotkov, E.V., Rudenko, V.M., 1999. Latent periodicity of protein sequences. J. Mol. Model. 5 (6), 103–115.

Kullback, S., 1997. Information Theory and Statistics. Courier Corporation, New York.

Louie, E., Ott, J., Majewski, J., 2003. Nucleotide frequency variation across human genes. Genome Res. 13 (12), 2594–2601.

McCarroll, S.A., Altshuler, D.M., 2007. Copy-number variation and association studies of human disease. Nat. Genet. 39, S37–S42.

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.

Smit, A.F., Hubley, R., Green, P., 1996. Repeatmasker Open-3.0.

Suvorova, Y.M., Korotkova, M.A., Korotkov, E.V., 2014. Comparative analysis of periodicity search methods in DNA sequences. Comput. Biol. Chem. 53, 43–48.

Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13 (1), 36–46.

Versalovic, J., Koeuth, T., Lupski, R., 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucl. Acids Res. 19 (24), 6823–6831.

Wang, J., Liu, G., Zhao, J., 2012. Some features of Fourier spectrum for symbolic sequences. Numer. Math. A J. Chin. Univ. 4 (24), 341–356.

Wang, J., Yin, C., 2016. A Fast Algorithm for Computing the Fourier Spectrum of a Fractional Period. arXiv preprint arXiv:1604.01589.

Yin, C., Wang, J., 2016. Periodic power spectrum with applications in detection of latent periodicities in DNA sequences. J. Math. Biol. 73 (5), 1053–1079.

Yin, C., Yau, S.S.-T., 2005. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. J. Comput. Biol. 12 (9), 1153–1165.

Yin, C., Yin, X., Wang, J., 2015. A novel method for comparative analysis of DNA sequences by Ramanujan–Fourier transform. J. Comput. Biol. 21 (12), 867–879.