Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 43596, 7 pages
doi:10.1155/2007/43596

Research Article

A Novel Signal Processing Measure to Identify Exact and Inexact Tandem Repeat Patterns in DNA Sequences

Ravi Gupta, Divya Sarthi, Ankush Mittal, and Kuldip Singh

Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, Roorkee 247 667, Uttaranchal, India

Received 6 September 2006; Revised 20 November 2006; Accepted 7 December 2006

Recommended by Yue Wang

The identification and analysis of repetitive patterns are active areas of biological and computational research. Tandem repeats in telomeres play a role in cancer and hypervariable trinucleotide tandem repeats are linked to over a dozen major neurodegenerative genetic disorders. In this paper, we present an algorithm to identify the exact and inexact repeat patterns in DNA sequences based on orthogonal exactly periodic subspace decomposition technique. Using the new measure our algorithm resolves the problems like whether the repeat pattern is of period P or its multiple (i.e., $2P$, $3P$, etc.), and several other problems that were present in previous signal-processing-based algorithms. We present an efficient algorithm of $O(NL_w \log L_w)$, where N is the length of DNA sequence and L_w is the window length, for identifying repeats. The algorithm operates in two stages. In the first stage, each nucleotide is analyzed separately for periodicity, and in the second stage, the periodic information of each nucleotide is combined together to identify the tandem repeats. Datasets having exact and inexact repeats were taken up for the experimental purpose. The experimental result shows the effectiveness of the approach.

1. INTRODUCTION

A direct or tandem repeat is the same pattern recurring on the same strand in the same nucleotide order, for example, TGAC recurs as TGAC. Tandem repeats play significant structural and functional roles in DNA. They occur in abundance in structural areas such as telomeres, centromeres, and histone binding regions [1]. They also play a regulatory role near genes and perhaps even within genes. Both degenerative diseases and cancer correlate to regions containing tandem repeats. Over a dozen of human degenerative diseases [2, 3], such as Huntington's disease, fragile X syndrome, mytonic dystrophy, and others, are associated with hypervariability of tandem repeats. Short tandem repeats are used as convenient tool for genetic profiling of individuals [4]. Thus, identification and analysis of repetitive DNA is an active area of biological and computational research.

The main objectives of repetitive pattern identification algorithms are to identify its periodicity, its pattern structure, its location and its copy number. The algorithmic challenges for repeat pattern identification problem are lack of prior knowledge regarding the composition of the repeat pattern and presence of inexact and hidden repeats. Inexact repeats are formed due to mutations of exact repeats and are thought to be representation of historical events associated with sequence. Thus, it is important for any repetitive pattern identification algorithm to identify inexact in addition to exact repeat structures in a DNA sequence.

In this paper, we have presented a novel SP-based approach for identifying exact and inexact tandem repeats in DNA sequences. In past, several algorithms and measures based on heuristic, combinatorial, dynamic programming, and SP approaches [5–13] have been proposed for finding tandem repeat structure in DNA sequences. SP-based algorithms for identifying tandem repeats have their own advantages because of its sensitivity towards detection of inexact repeats and application of faster signal processing tool like DFT. These algorithms also provide an easy solution to biologist or noncomputer experts because unlike non-SP algorithms which require a number of error tolerances parameters like match, edit distance, Hamming distance, and several other parameters which are very difficult to understand for any normal user, the SP-based algorithms require mainly one parameter which acts as a threshold for identifying repeats.

Previous SP solutions to repeat pattern identification problem include the application of discrete Fourier transform (DFT) [11, 12] and the application of short-time periodicity transform (STPT) [13]. In [11], DFT is used as a preprocessing tool for identifying the significant periodic regions through a sliding window analysis, and then an exact search method is used for finding the repetitive units. In [12], instead of a product spectrum a sum spectrum was proposed as a measure for identifying repeats. The product spectrum is especially sensitive to the presence of inexact repeats. An STPT-based approach for finding tandem repeats in DNA sequence is presented in [13]. Both DFT- and STPT-based techniques suffer from one major disadvantage while detecting inexact repeats. They cannot tell whether a repeat is of period P or its multiple, that is, $2P$, $3P$, and so on. In addition to this, the STPT-based algorithm has several other drawbacks which are discussed in the later section of this paper.

The contribution of this paper is in providing a novel SP application in the area of DNA sequence analysis. An exactly periodic subspace decomposition (EPSD) [14] based measure for identifying repeats is presented in this paper. EPSD technique, unlike the Fourier transform, is obtained by taking projection onto exactly periodic orthogonal multidimensional subspaces. By having subspaces of dimensions larger than one, the exactly periodic subspace (EPS) can better capture, in one coefficient, the periodic energy than the Fourier transform. Hence, the new measure of the algorithm is more sensitive than previous techniques for identifying repeats.

In addition to identification of exact repeats, the proposed measure is useful in identifying inexact and other hidden repeat patterns unannotated by GenBank database. The EPSD-based approach also helps in identifying whether a particular pattern is due to period P or its multiple. Thus the ambiguity that is present in [11–13] is taken care by our algorithm. The algorithm proposed in this paper first analyzes four nucleotide sequences separately and later on the results obtained are processed together to locate the tandem repeats. The algorithm presented runs in $O(NL_w \log L_w)$, where N is the length of the DNA sequence and L_W is the length of the window. Experiments were performed on various types of data sets. The data sets include the genes of degenerative disease having long exact tandem repeat; inexact, complex, and hidden repeats. Comparison with other techniques shows the effectiveness of our approach.

The paper is organized as follows. Section 2 initially provides a mathematical formulation of repeat pattern identification problem and later on briefly describes the EPSD technique. Section 3 presents a repeat pattern detection algorithm for identifying various repeat patterns present in the DNA sequence. In Section 4, the algorithm is applied on some actual DNA sequence and experimental result is presented. Conclusion and future work follow in Section 5.

2. MATHEMATICAL FORMULATION OF TANDEM REPEAT PATTERN IDENTIFICATION

The standard representation of genomic information by sequences of nucleotide symbols in DNA, RNA, or amino acids limits the processing of genomic information to pattern matching and statistical analysis. Providing mathematical representation to symbolic DNA sequences opens the possibility to apply signal processing techniques for the analysis of genomic data [15] and reveals features of genomes that would be difficult to obtain by using standard statistical and pattern matching techniques. The arbitrary assignment of a number to each symbol would impose a mathematical structure not present in the original data. Thus, a nucleotide mapping should be chosen such that it preserves the biological features and does not introduce any artifact into the mapped signal. For our algorithm, we have selected binary indicator sequence [16] representation for the DNA sequence. This mapping helps in formulating the tandem repeat identification problem analogous to period detection in signal processing.

2.1. *Numerical representation of DNA sequences*

Consider a DNA sequence $S[n] = s_1 s_2 \cdots s_L$ of length L, consisting of a sequence of a series of four nucleotides symbols $\{$A,C,G,T$\}$. The binary indicator sequences are obtained as follows:

$$S_\Omega[n] = \begin{cases} 1, & \text{if } S[n] = \Omega \text{ where } \Omega \in \Sigma(= \{\text{A,C,G,T}\}), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

2.2. *Definitions of different repeats in DNA sequences*

Definition 1. A subsequence $S'[n] = s_i s_{i+1} \cdots s_{i+l-1}$ of $S[n]$ is an exact tandem repeat (ETR) of period "p" and repeat pattern $\alpha = r_1 r_2 \cdots r_p$ (where "i" is the starting position and "l" is the length of ETR), if the following conditions are satisfied.

(1) $\lfloor l/p \rfloor \geq 2$, *where* $\lfloor l/p \rfloor$ is the count for pattern (α), that is, number of times α has occurred in subsequence $S'[n]$. The count of repeat pattern (α) should at least be equal to two.
(2) $\Lambda = \{r_1, r_2, \ldots, r_p\}$, where $\Lambda \subseteq \Sigma$ and $|\Lambda| \geq 1$.
(3) $S_\Delta[n]$ is p-periodic for all $\Delta \in \Lambda$, where $i \leq n \leq i+l$.

For example, if $S[n]$ = GGCATACT**ACGACGACG**CCG, then $S'[n]$ = ACGACGACG, $i = 9$, $p = 3$, $l = 9$, $\lfloor l/p \rfloor = 3$, α = ACG, $\Lambda \equiv \{$A,C,G$\}$, and $S_\text{A}[n]$, $S_\text{C}[n]$, $S_\text{G}[n]$ are 3-periodic sequence.

Definition 2. A subsequence $S''[n] = s_i s_{i+1} \cdots s_{i+l-1}$ of $S[n]$ is an inexact tandem repeat (InTR) of period "p" and consensus repeat pattern $\alpha = r_1 r_2 \cdots r_p$ (where "i" is the starting position and "l" is the length of InTR), if the following conditions are satisfied.

(1) $\lfloor l/p \rfloor \geq 2$.
(2) $\Lambda = \{r_1, r_2, \ldots, r_p\}$, where $\Lambda \subseteq \Sigma$ and $|\Lambda| \geq 1$.
(3) $S_\Delta[n]$ is nonperiodic, for at least one $\Delta \in \Lambda$, *where* $i \leq n \leq i + l$.
(4) For all $\Delta \in \Lambda$, p-period measure of $S_\Delta[n] \geq$ threshold.

For example, if $S[n]$ = GGCATACACAGACACGCCGGCG, then $S''[n]$ = ATACACAGACAC, $i = 4$, $p = 2$, $l = 12$, α = AC, $\Lambda \equiv \{$A,C$\}$, and $S_\text{A}[n]$ is 2-periodic sequence (not necessarily exact).

From the above formulation, we notice that the repeat identification in DNA is analogous to period detection in signals. So, the knowledge of periodicity in the binary signals (i.e., $S_\Omega[n]$) helps in identifying tandem repeats in the DNA sequence. Thus, the main objective of SP algorithm for this problem is to develop a good measure for identifying periods in the binary signals.

In [11], Sharma et al. proposed a DFT-based algorithm (SRF) for identifying tandem repeats in DNA sequence based on sum spectra. The sum spectra measure is obtained by summing up the spectra of each binary subsequence. However, in case of InTR, not all the binary subsequences are exactly periodic, and hence the sum spectra measure is not effective when InTR are to be identified in DNA sequences. Also, it cannot tell whether the repeat pattern is of period P, $2P$, or its multiple.

A STPT-based periodicity explorer (PE) algorithm is proposed in [13] for identifying tandem repeat. The PE algorithm has several shortcomings. The nucleotide mapping in [13] was taken as follows: $\mathrm{A} = 1+j$, $\mathrm{C} = -1+j$, $\mathrm{G} = -1-j$, and $\mathrm{T} = 1-j$, where $j = \sqrt{-1}$. Let the two DNA sequences be ACATACAC and ACAGACAC. The projection of the DNA sequences onto the periodic subspace P_2 (where P is the set of all periodic sequences) is given by $\{(1+j), (-0.5+0.5j), (1+j), (-0.5+0.5j), (1+j), (-0.5+0.5j), (1+j), (-0.5+0.5j)\}$ and $\{(1+j), (-1+0.5j), (1+j), (-1+0.5j), (1+j), (-1+0.5j), (1+j), (-1+0.5j)\}$, respectively. And the periodogram coefficient values for the DNA sequence for projection on P_2 subspace are 0.75 and 0.895, respectively. By comparing the two DNA sequences, we observe that even though the two DNA sequences have equal degree of period 2 component (differ just by one symbol from becoming ETR), the projection of DNA sequences are different and also the periodogram coefficient obtained are different. This shows that the periodogram coefficient cannot act a good estimator for measuring periodicity.

The PE algorithm is designed to be executed separately for every period because the periodicity transform provides nonorthogonal decomposition of the signal. This means that the run time of the PE algorithm is $O(NWP_{\max})$, where N is the length of analyzed DNA sequence, W is the window size, and $P_{\max}$ is the maximum period. Also, like STPT, it cannot tell whether the tandem repeat present in the DNA sequence is of period P or multiple of P (i.e., $2P$, $3P$, etc.). Thus, we need an SP algorithm which can take care of the shortcomings present in previous approaches for identifying different types of repeat present in DNA sequences. In the algorithm proposed later on in this paper, a novel signal processing measure based on EPSD [14] technique is provided for identifying ETR and InTR in DNA sequence and overcomes the shortcomings in previous algorithms.

2.3. *Exactly periodic subspace decomposition*

The exactly periodic subspace decomposition (EPSD) technique was proposed by Muresan and Parks [14]. The EPSD technique generates orthogonal subspaces that correspond to periods ranging from 1 up to the maximum expected subperiod of the input signal S. The energy of the expected subperiods is obtained by taking orthogonal projections of S onto these different orthogonal subspaces. The key idea behind the EPSD technique is the concept of exactly periodic signals (EPS). The definition of exactly periodic signal is given as follows.

Definition 3. A signal S is of exactly period P if S is in Φ_P (where Φ_P is the subspace of the signal of period P) and the projection of S onto subspace $\Phi_{P'}$ for all $P' < P$ (where $\Phi_{P'}$ is the subspace of signal of period P') [14].

Thus, a signal of exactly period P is not exactly period $2P$, $3P$, and so forth, although it continues to be of period $2P$, $3P$, and so forth. Also, not every periodic signal is exactly periodic, but every exactly periodic signal is periodic. Some of the important properties of the EPSD technique are the following.

(1) The EPSD technique completely decomposes the input signal $S \in \mathbb{R}^n$ into exactly periodic orthogonal components corresponding to each of the exactly periodic signals of n and all possible factors of n.

(2) Unlike the STPT [13], the decomposition of the EPSD technique is unique. Thus, the input signal can be uniquely decomposed on the orthogonal subspaces.

(3) The EPSD of signal is achieved by taking projections onto exactly periodic orthogonal multidimensional subspaces of periods that divides n, whereas the discrete Fourier transform is obtained by taking orthogonal projections onto one-dimensional (1D) complex exponentials $e^{j((2\pi)/N)k}$ with frequencies (k/N), $k = 0,\dots,N-1$. The EPS is spanned by a collection of Fourier exponentials, which is dictated by the period. Thus, by having spaces of dimensions larger than one, EPS can capture in one coefficient the periodic energy better than the Fourier transform.

In [14], the EPSD technique was proposed to identify periodic signal by considering the entire input signal, that is, it provides information about the periods that are present in complete input data sequence. However, in tandem repeat identification problem, even though the core objective is to identify periods in DNA sequences, there is one major difference. Instead of looking for periods that are present in entire input DNA sequence, we have to look for local periodic information because most of the tandem repeats that are present in the DNA sequences are localized to small portion of the complete genome. In addition, the tandem repeats forms only small fraction of total genome. Thus, the main objective of tandem repeat identification program is to provide the localized periodic information. We have adapted the EPSD technique for our problem to provide a measure for localized periodic information that is present in the mapped DNA sequences.

Instead of analyzing the complete input DNA sequence in one go, we divide the DNA sequence into a set of subsequences defined by a pointwise multiplication of the original DNA sequence by a stationary window. The EPSD technique is then applied to the resulting subsequences. Let the window be represented by W_i of length L_w and beginning at ith

```
(1) Accept window size ($L_w$), maximum period ($P_{\max}$)
(2) for $i = 1$ to $N + L_w - 1$ do // $N$ is the length of DNA sequence
(3) $S_{W,i}[n] = S_{W,i}[n] - \overline{S}_{W,i}[n]$, where $\overline{S}_{W,i}[n] = \text{MEAN}(S_{W,i}[n])$
(4) $\alpha_{w,i}[1,\dots,P_{\max}] = \text{EPSD}(S_{W,i}[n], P_{\max})$
(5) $\pi_{W,i}[1,\dots,P_{\max}] = \dfrac{\|\alpha_{W,i}[1,\dots,P_{\max}]\|^2}{\|S_{W,i}[n]\|^2}$
(6) OUTPUT($p_i, \pi_{W,i}[p_i]$), where $\pi_{W,i}[p_i] \leftarrow \max(\pi_{W,i}[1],\dots,\pi_{W,i}[P_{\max}])$
```

Algorithm 1: Calculation of repeat coefficient for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$.

element, where

$$W_i[n] = \begin{cases} 1, & n = i, i+1, \dots, i + L_w - 1, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

The localized portion of the sequence S, $S_{W,i}$ is defined as

$$S_{W,i}[n] = S[n] \cdot W_i[n]. \tag{3}$$

3. TANDEM REPEAT DETECTION ALGORITHM

The objectives of our proposed algorithm are to identify the position, period, and the length of repeat patterns in DNA sequences. For identifying repeats, the symbolic DNA sequences are first mapped into four digital signals and then EPSD mathematical tool is applied. Later on, repeat coefficient measure is calculated for each window and the potential repetitive patterns are reported depending on the value of input parameters provided by the user. The algorithm is designed to identify tandem repeats from period 2 to maximum period ($P_{\max}$) provided by the user within an observation window of size L_w. The complete repeat detection process is divided into three major steps. We describe next our proposed algorithm.

Step 1 (nucleotide mapping of DNA sequence $S[n]$ into four nucleotide subsequences). The nucleotide mapping procedure was discussed in the previous section. In this step, we obtain four binary subsequences ($S_A[n]$, $S_C[n]$, $S_G[n]$, and $S_T[n]$) using (1) that act as input signals for our algorithm.

Step 2 (calculation of tandem repeat coefficient for subsequences). For identifying the position of the tandem repeats in DNA sequences, we use a sliding window-based approach. The algorithm for calculating period with maximum energy for the input DNA sequence of length N and input parameters ($P_{\max}, L_w$) is provided (see Algorithm 1), where the value of $P_{\max}$ can vary from 2 to $L_w/2$. The prior knowledge of maximum repeat pattern size restrict our search to pattern size $P_{\max}$. However, if the user does not have prior knowledge, then the value of $P_{\max}$ can be fixed to $L_w/2$. In step (3) of the algorithm, we remove the dc component (i.e., period-1) from the input signal. This step helps in removing the repeats that due to single base repeat pattern, for instance, repeat like AAAAA in DNA sequence ACGACAAAAACAACG because the repeat pattern of period 1 is of no interest. In step (4), the energy of the input signal is decomposed on the subspaces from 2 to $P_{\max}$ using EPSD technique. The energies of the subspaces are stored in the array $\alpha_{w,i}$. The array $\pi_{W,i}$, which is calculated in step (5), measures the fraction of power of the periodic subspaces from 2 to $P_{\max}$. The value $\pi_{W,i}$ acts as an indicator for identifying the local periodicities of the input sequence and is said as *tandem repeat coefficient*. And finally in step (6), we obtain a tuple $\langle p, \pi_{W,i}[p] \rangle$ for each window where p is the periodic subspace that have maximum fraction of power in the subsequence for the window positioned at i. Algorithm 1 unlike the PE algorithm needs just a single scan for identifying the period ($\le P_{\max}$) of repeat patterns in the input DNA sequence. This step is performed on all four binary subsequences obtained from the previous step.

Step 3 (identification and characterization repeat from binary subsequences). In this step, we first identify the repeats that are present in all four binary subsequences utilizing the value of threshold parameter (τ) provided by the user and tuple $\langle p_i, \pi_{W,i}[p_i] \rangle$ calculated in the previous step using EPSD technique. A repeat is represented by tuple $\langle \Omega, i, l, p \rangle$, *where* $\Omega \in \{\text{A, C, G, T}\}$, i is the starting position of the repeat (position of the window), l is the length of the repeat, and p is the period of repeat. A repeat satisfies the following conditions:

(i) $\pi_{W,i}, \pi_{W,i+1}, \dots, \pi_{W,i+l-1} \ge \tau$ (threshold);
(ii) $p_i = p_{i+1} = \cdots = p_{i+l-1} = p$.

After the repeats in each subsequences are identified, we process all four subsequences together and classify the repeats into ETR and InTR based on the definitions provided in previous section.

4. EXPERIMENTAL RESULTS

To demonstrate the capabilities of the repeat pattern identification algorithm, experiments were performed on datasets of some actual DNA sequences available at GenBank database. The proposed algorithm was implemented in Matlab 7.0 for Microsoft Windows ® platform. The EPSD function was implemented using the code available at http://dsplab.ece.cornell.edu/about/about_software.htm for noncommercial use. The datasets were selected such that the experiment covers exact and inexact (complex, dispersed, and hidden) repeat patterns. Some of the typical results are provided in this section. We also provide results obtained from other tandem repeat identification algorithm when applied to the DNA sequences considered for analysis.

DATASET 1

Myotonic dystrophy disease, the most common muscular dystrophy in humans, is caused by an expansion of the CTG

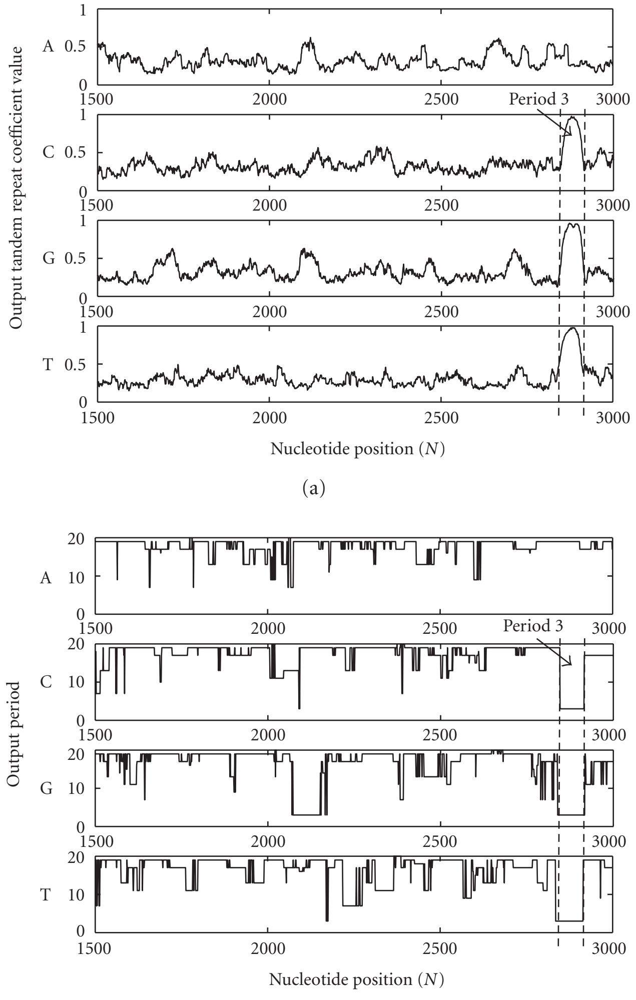


Figure 1: (a) The tandem repeat coefficient value of subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ and (b) the output period obtained for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for DNA sequence (Accession: XM_027572, length = 3436 base pair (bp)) with input parameters (window length = 80 and maximum period = 20).

repeat located in the 3′-UTR (untranslated region) of dystrophia myotonica protein kinase (DMPK) gene [17]. The 3′-UTR region is present after a coding region in a DNA sequence. For a normal person, the repeat number of CTG is less than 35 and for a person suffering from myotonic dystrophy the CTG count is above 50 [3]. This dataset consists of DNA sequence (GenBank: XM_027572, length = 3436 base pairs (bp)) of Homo sapiens DMPK gene sequenced under NCBI annotation project.

The DNA sequence is tested with input parameters for window size (L_w) = 40 and maximum period ($P_{\max}$) = 10 and threshold (τ) = 0.95. The tandem repeat coefficients obtained for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ are shown in Figures 1(a) and 1(b); we provide the output period obtained for the subsequences. The subsequences $S_C[n]$, $S_G[n]$, and $S_T[n]$ have repeat coefficient value greater than 0.95 from 2876 to 2967 and the corresponding output period is 3 (shown in Figure 1(b)). An exact trinucleotide tandem repeat pattern CTG of repeat length 62 (repeat number ≈ 21), beginning at 2890, was identified in the DNA sequence. The protein coding sequence for human DMPK gene is 779–2668 bp. And as the identified tandem repeat lies after 2668 bp in DMPK gene sequence, this confirms the presence of CTG repeat in 3′-UTR of human DMPK. Apart from exact tandem repeats, weak patterns of period 3 were identified for nucleotides C (beginning at 1864, length of 21) and G (beginning at 2114, length of 63).

Table 1: Repeat patterns identified in HSVDJSAT DNA sequence.

Program	Consensus period	Repeat region
Our algorithm	$2^{(a),(c)}$	825–865
	$9^{(a),(c)}$, $10^{(a),(c)}$, $19^{(b),(d)}$, $49^{(b),(d)}$	1177–1545
Hauth program	9, 10, 19, 37, 38, 48	1197–1538
TRF 4.0$^{(e)}$	$2^{(c)}$	826–856
	$10^{(c)}$	1199–1539
	$19^{(d)}$	1190–1539
	$49^{(d)}$	1195–1539

$^{(a)}$Maximum period size ($P_{\max}$) ≤ 10, $^{(b)}$Maximum period size ($P_{\max}$) > 10. $^{(c)}$Simple tandem repeat, $^{(d)}$Multiperiod tandem repeat.

$^{(e)}$Alignment parameter (match, mismatch, indel) = (2, 7, 7), minimum alignment score = 30, and maximum period size = 50.

Experiment was also conducted using TRF 4.0 and PE for a maximum period size equal to 10. TRF 4.0 with default input parameters provides output consisting of tandem repeat of pattern TGC starting at 2890 and repeat length 62. The PE program provided output pattern of period 3 (TGC), period 6 (TGCTGC), and period 9 (TGCTGCTGC).

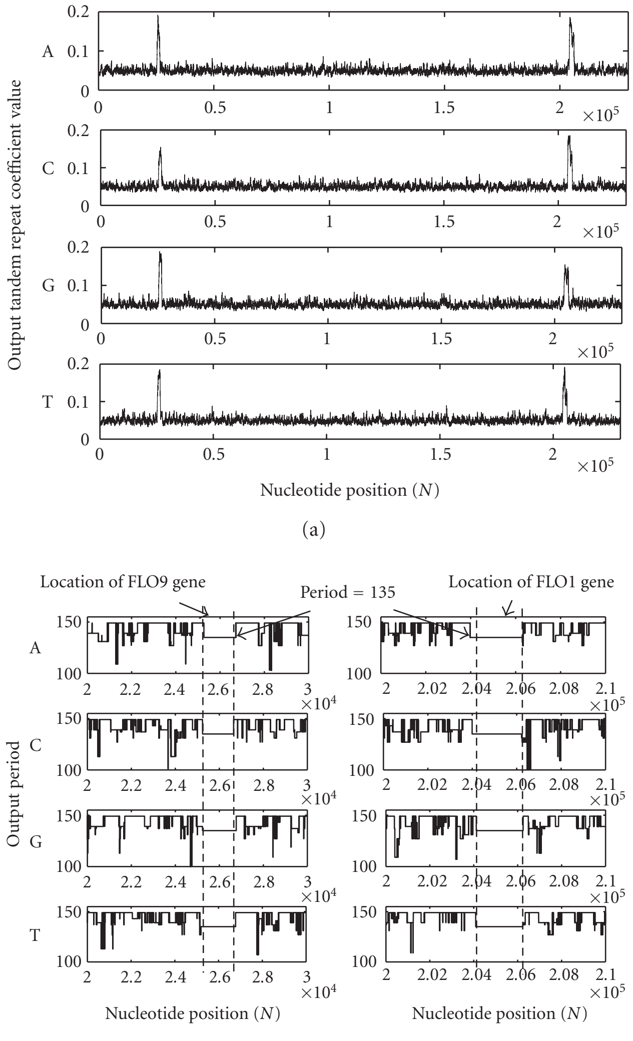
DATASET 2

The analysis of Homo sapiens, GeneBank Locus: HSVDJSAT of length 1985 bp, is provided in this example. This DNA sequence consists of simple and multiperiod tandem repeat patterns. Periods of size 2, 9, 10, 19, and 48 were identified in the DNA sequence. The details regarding the identified repeats are provided in Table 1. The consensus tandem repeat patterns of size 2, 19, and 49 reported by our algorithm are: AC, CTGGGAGAGGCTGGGATTG, CTGGGAGAGGCTGGGAGAG, GAGGCTGGGAGAGGCTGGGAGAG∗CTGGGAGAGGCTG∗GATTGCTGGGA (where ∗ represents any of the four nucleotides, i.e., A, C, G, or T). Tests were also performed by tandem repeat finder (TRF) 4.0 [5, 18] and Hauth program [10] for identifying repeats. In [19], Hauth reported the 49 period as period of 48 and missed the simple repeat pattern of period 2. The TRF 4.0 program missed the tandem repeat pattern of period size 9.

DATASET 3

The complete chromosome I sequence contains two flocculation genes (FLO1 and FLO9), one at each end of the chromosome, that each contains a tandem repeat region having similar 135 bp pattern [20]. The GeneBank details of the DNA sequence and genes (FLO1 and FLO9) are as follows:

locus: NC_001133, total base pairs: 230208;



(a)

(b)

Figure 2: (a) The tandem repeat coefficient value of subsequences $S_{\mathrm{A}}[n]$, $S_{\mathrm{C}}[n]$, $S_{\mathrm{G}}[n]$, $S_{\mathrm{T}}[n]$ and (b) the output period obtained for subsequences $S_{\mathrm{A}}[n]$, $S_{\mathrm{C}}[n]$, $S_{\mathrm{G}}[n]$, $S_{\mathrm{T}}[n]$ for DNA sequence (Accession: NM_001133, length = 230208 bp) with input parameters (window length = 600 and maximum period = 150).

organism: Saccharomyces cerevisiae (baker's yeast);
gene: FLO1, region in DNA sequence: 24001–27969;
gene: FLO9, region in DNA sequence: 203394–208007.

The DNA sequence is processed by the algorithm with input parameters, window size $(L_w) = 600$ and maximum period $(P_{\max}) = 150$. The outputs (i.e., repeat coefficients and maximum period) of the algorithm for the nucleotide subsequences are provided in Figures 2(a) and 2(b). Two sharp peaks are present in Figure 2(a). These peaks are due to presence of strong tandem repeats in the DNA sequence at these positions. The first peak starts at 25 324 and lasts for 1842 bp. The maximum period for this region as shown in Figure 2(b) is 135. This tandem repeat region lies in gene FPO9. The second peak starts at 204 207 and lasts for 2466 bp. This region also has maximum period of 135 bp. However, the total number of copies for this tandem repeat is higher than the previous one. The result confirms the presence of strong tandem repeats which are present in FLO1 and FLO9 genes of saccharomyces cerevisiae, chromosome I.

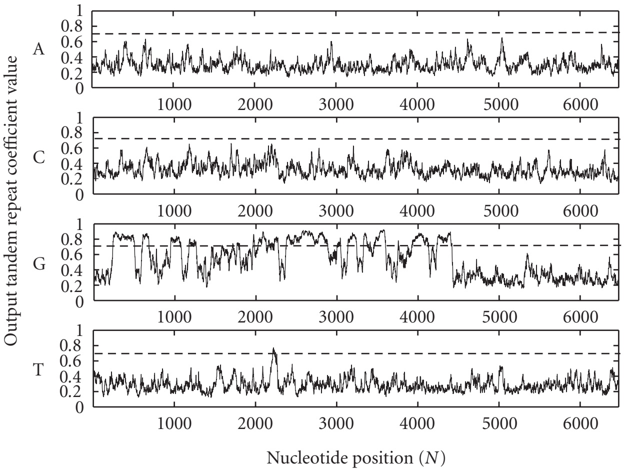


Figure 3: Tandem repeat coefficient value of subsequences $S_{\mathrm{A}}[n]$, $S_{\mathrm{C}}[n]$, $S_{\mathrm{G}}[n]$, $S_{\mathrm{T}}[n]$ for DNA sequence (Accession: NM_001847, length = 6574 bp) with input parameters (window length = 100 and maximum period = 20).

DATASET 4

The analysis of Homo sapiens collagen gene, GenBank accession no. NM_001847 of length 6574 bp containing weak tandem repeat pattern is provided in this example. The tandem repeat coefficient obtained for subsequences $S_{\mathrm{A}}[n]$, $S_{\mathrm{C}}[n]$, $S_{\mathrm{G}}[n]$, $S_{\mathrm{T}}[n]$ for window size $(L_w) = 100$ and maximum period $(P_{\max}) = 20$ is shown in Figure 3. In the figure, subsequence $S_{\mathrm{G}}[n]$ has significant repeat coefficient value from 250 to 4400, while for subsequence $S_{\mathrm{T}}[n]$ the repeat coefficient is above (threshold = 0.7) from 2233 to 2326. However, for other subsequences, that is, $S_{\mathrm{A}}[n]$ and $S_{\mathrm{C}}[n]$, the value of repeat coefficient lies between 0.4 and 0.6. This shows the presence of repetitive pattern involving nucleotide G and T.

Tests were also performed using PE and TRF program. PE program gave tandem repeat of period 9 and multiple of 9 (i.e., 18, 27, etc.). This is due to problem with the PE algorithm because it cannot distinguish whether a repeat is of period p or its multiple. However, this problem did not appear in our algorithm because of unique decomposition property of EPSD technique. The TRF program provided two tandem repeat region of period 9 starting at 963 and 1404. Both PE and TRF fail to inform the user regarding hidden periodicity of nucleotide G. This has happened because the TRF and PE programs are designed only to detect tandem repeat and not hidden periodicity of individual nucleotides in DNA sequences.

DATASET 5

In our last dataset, a human microsatellite repeat (GenBank Accession: M65145) is taken up for analysis. Figure 4 shows the periods identified in the DNA sequence. It is clear that the DNA sequence contains two repeat regions of period 2 and 11. The dinucleotide repeats of pattern TG occur

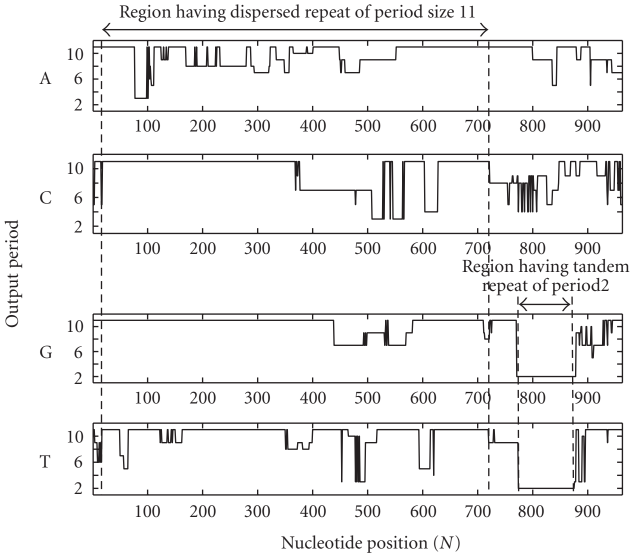


Figure 4: Output period of subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for DNA sequence M65145 with input parameters (window length = 110 and maximum period = 11).

between positions 780 and 933 bp (GenBank annotation is between 860 and 900 bp). And the 11-mer repeats are located between 92 and 781 bp (unannotated by GenBank). The analysis of the 11-mer repeat region of the DNA sequence reveals the dispersed (hidden repeat) copy of the 11-mer TGACTTTGGGG. The TRF program was unable to detect the 11-mer repeats in the DNA sequence. This clearly shows the advantage of our algorithm in locating dispersed or hidden periodic patterns.

5. CONCLUSION

A novel SP-based approach is presented in this work. It has the potential to identify and locate exact and inexact repeat pattern in DNA sequences. A new measure based on EPSD technique is proposed in this paper. A DNA sequence is converted into a digital subsequences and repeat coefficient measure is computed. The algorithm is designed to analyze each nucleotide sequence separately, and later on result of individual nucleotides are combine together to report repeats. The algorithm runs in $O(NL_w \log L_w)$ and is computationally faster than PE algorithm which runs in $O(NL_w P_{\max})$, where N is the length of the analyzed DNA sequence, L_w is the window size, and $P_{\max}$ is the maximum period to be identified. Our algorithm also resolves the problems like whether the repeat pattern is of period P or its multiple (i.e., $2P$, $3P$, etc.) and other issues related to detection of inexact tandem repeats that were present in previous signal-processing-based algorithms. The experimental results and comparison with other algorithms show the effectiveness of our algorithm. Design of automatic selection of window size for different repeat period can be taken up for future work.

REFERENCES

[1] W. C. Hahn, "Telomerase and cancer: where and when?" *Clinical Cancer Research*, vol. 7, no. 10, pp. 2953–2954, 2001.

[2] R. R. Sinden, V. N. Potaman, E. A. Oussatcheva, C. E. Pearson, Y. L. Lyubchenko, and L. S. Shlyakhtenko, "Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA," *Journal of Biosciences*, vol. 27, no. 1, supplement 1, pp. 53–65, 2002.

[3] E. Y. Siyanova and S. M. Mirkin, "Expansion of trinucleotide repeats," *Molecular Biology*, vol. 35, no. 2, pp. 168–182, 2001.

[4] K. Tamaki and A. J. Jeffreys, "Human tandem repeat sequences in forensic DNA typing," *Legal Medicine*, vol. 7, no. 4, pp. 244–250, 2005.

[5] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences," *Nucleic Acids Research*, vol. 27, no. 2, pp. 573–580, 1999.

[6] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich, "REPuter: the manifold applications of repeat analysis on a genomic scale," *Nucleic Acids Research*, vol. 29, no. 22, pp. 4633–4642, 2001.

[7] R. Kolpakov, G. Bana, and G. Kucherov, "mreps: efficient and flexible detection of tandem repeats in DNA," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3672–3678, 2003.

[8] G. M. Landau, J. P. Schmidt, and D. Sokol, "An algorithm for approximate tandem repeats," *Journal of Computational Biology*, vol. 8, no. 1, pp. 1–18, 2001.

[9] E. F. Adebiyi, T. Jiang, and M. Kaufmann, "An efficient algorithm for finding short approximate non-tandem repeats," *Bioinformatics*, vol. 17, supplement 1, pp. S5–S12, 2001.

[10] A. M. Hauth and D. A. Joseph, "Beyond tandem repeats: complex pattern structures and distant regions of similarity," *Bioinformatics*, vol. 18, supplement 1, pp. S31–S37, 2002.

[11] D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy, "Spectral repeat finders (SRF): identification of repetitive sequences using Fourier transformation," *Bioinformatics*, vol. 20, no. 9, pp. 1405–1412, 2004.

[12] T. T. Tran, V. A. Emanuele II, and G. T. Zhou, "Techniques for detecting approximate tandem repeats in DNA," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 449–452, Montreal, Quebec, Canada, May 2004.

[13] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2280–2287, 2003.

[14] D. D. Muresan and T. W. Parks, "Orthogonal, exactly periodic subspace decomposition," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2270–2279, 2003.

[15] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[16] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Computer Applications in the Biosciences*, vol. 13, no. 3, pp. 263–270, 1997.

[17] A. D. Otten and S. J. Tapscott, "Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 12, pp. 5465–5469, 1995.

[18] G. Benson, "Tandem Repeat Finder," http://tandem.bu.edu/trf/trf.html.

[19] A. M. Hauth, "Identification of tandem repeats simple and complex pattern structures in DNA," Ph.D. dissertation, University of Wisconsin-Madison, Madison, Wis, USA, 2002.

[20] H. Bussey, D. B. Kaback, W. Zhong, et al., "The nucleotide sequence of chromosome I from Saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 9, pp. 3809–3813, 1995.