14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

EXACTLY PERIODIC SUBSPACE DECOMPOSITION BASED APPROACH FOR IDENTIFYING TANDEM REPEATS IN DNA SEQUENCES

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

Department of Electronics & Computer Engineering, Indian Institute of Technology Roorkee
Roorkee, Uttaranchal, 247 667, India
phone: + (91) 9897043997, email: {rgcsedec, samaypec, ankumfec, ksd56fec}@iitr.ernet.in

ABSTRACT

The identification and analysis of tandem repeats is an active area of biological and computational research. Tandem repetitive structures in telomeres plays a role in cancer and hyper-variable, trinucleotide tandem repeats are linked to over a dozen major degenerative diseases. They also play a very crucial role in DNA fingerprinting.

In this paper we present an algorithm to identify the exact and inexact tandem repeats in DNA sequences based on orthogonal exactly periodic subspace decomposition technique. The algorithm uses a sliding window approach to identify the location of tandem repeats and other patterns that are present in DNA sequence due to repetition of individual nucleotides. Our algorithm also resolves the problems that were present in periodicity explorer algorithm for identifying tandem repeats. The time complexity of the algorithm, when searching for repeats for window size 'W' in a DNA sequence S of length N is O(NWlgW). We present some experimental results concerning to sensitivity of our algorithm.

1. INTRODUCTION

One of the significant genomic achievements in recent times has been the development of fast methods for sequencing and proteins. This has enabled the creation of large databases which can be processed by considering sequences of nucleic acids (DNA, RNA) and amino acids (proteins) as string of characters. This type of processing is one of the fundamental basis of modern bioinformatics and has transformed biology from a laboratory science to a computational one.

For the last few decades, the major thrust of DNA and protein analysis has been on string matching, either with the goal of obtaining a precise solution with dynamic programming or some heuristic techniques for obtaining a faster solution. However, these heuristic methods do not work well on repetitive structures.

A repeat is a recurrence of a pattern. A DNA pattern recurs in four ways: direct, indirect complement of reverse complement. In DNA, most repetitions occur as tandem or reverse complement repeats. A tandem repeat is a string that can be divided into identical substrings, e.g., ACGACGACG. In eukaryotic genomes, tandem repeats are involved in various regulation mechanisms. Tandem repeats are also involved in human neurological disorders, such as huntington's disease, fragile X syndrome, mytonic dystrophy and others [1, 2]. A major application of short tandem repeats is based on the inter-individual variability in copy number of certain repeats occurring in single loci. This feature makes tandem repeats a convenient tool for genetic profiling of individuals [3, 4]. Thus, it is critical to both the assembly and analysis of genomic sequences to identify and characterize tandem repeat sequence.

Previous signal processing techniques for the identification of tandem repeats in DNA sequences include the application of discrete Fourier transform (DFT) [5, 6] and the application of short-time periodicity transform [7]. In [5] the DFT is used as a pre-processing tool for identifying the significant periodic regions through a sliding window analysis, and later on an exact search method is then used for finding the repetitive units. In [6] a product spectrum instead of sum spectrum was proposed as a measure for identifying repeats. The product spectrum is especially sensitive to the presence of inexact repeats. A short-time periodicity transform based approach for finding tandem repeats in DNA sequence is presented in [7]. This technique is useful only for exact tandem repeats and the inexact tandem repeats which are due to substitution of nucleotides in the DNA sequence.

Apart from signal processing techniques several other algorithms [8, 9, 10, 11, 12, 13] have been proposed for detecting exact and inexact repeats in DNA sequences. Each of the algorithms has its own limitations and assumptions. In [8], the period of the repeats is limited to 2000 base pairs (bp) as of version 4.0. In [9] there are practical memory constrains resulting from the pattern extension algorithm. The algorithm [11] has no limitations on period size but does not deal with insertion or deletions (indels) directly.

In this paper, we present exactly periodic subspace decomposition (EPSD) [14] based approach for identifying the tandem repeats in the DNA sequences. The EPSD technique unlike periodicity transform [15] gives a unique decomposition of the signal on the periodic subspace and hence finds energy of each periodic subspace without any ambiguity. The algorithm presented runs in O($NW\lg W$) where N is the length of the DNA sequence and W is the length of the window.

The paper is organized as follows. Section II describes about the exactly periodic subspace decomposition. Section III presents a tandem repeat detection algorithm for identifying

both strong and weaker tandem repeat present in the DNA sequence. In Section IV the algorithm is applied on some actual DNA sequence and experimental result is presented. Conclusion and future work follow in Section V.

2. EXACTLY PERIODIC SUBSPACE DECOMPOSITION

Definition 1: A signal S is of exactly period P if S is in $R(\psi^P)$, and the projection of S onto $R(\psi^{\overline{P}})$ is zero for all $\overline{P} < P$ (where $R(\psi^{\overline{P}})$ is the subspace of signal of period $\overline{P}$).

With the above definition, a signal of exactly period P is not exactly period $2P$, $3P$, etc. In addition, not every periodic signal is exactly periodic, but every exactly periodic signal is periodic. For example, an exactly periodic 4 signal is

$$R = [1, 1, -1, -1, 1, 1, -1, -1, 1, 1, -1, -1]$$

The exactly periodic subspace decomposition technique finds the subspace corresponding to the signal of exactly periodic P and shows that these subspaces are orthogonal to each other. The study of this problem in time domain begins with the following definition.

Definition 2: Let $p_1, p_2, \ldots, p_m$ are the divisors of P, then defining $\psi^P_{p_1,\ldots p_m}$ be the matrix whose range is the orthogonal complement of $R[\psi^{p_1} \ldots \psi^{p_m}]$ inside $R(\psi^P)$:

$$R(\psi^P_{p_1\ldots p_m}) = R(\psi^P) \cap (R[\psi^{p_1} \ldots \psi^{p_m}]) \perp \qquad (1)$$

If p_i includes 1 in it then $R(\psi^P_{p_1\ldots p_m})$ is the subspace corresponding to signal of exactly period P. The following lemma proves the definition introduced above.

Lemma 1: Given a signal of length L (L is multiple of P_1 and P_2), let $R(\psi^{P_1})$ and $R(\psi^{P_2})$ be the subspaces corresponding to period P_1 and P_2. In addition $R(\psi^{P_3})$ be the subspace corresponding to period P_3, where P_3 is the greatest common divisor of P_1 and P_2. Then $R(\psi^{P_3})$ is the intersection of $R(\psi^{P_1})$ and $R(\psi^{P_2})$. Moreover, the orthogonal complement of $R(\psi^{P_3})$ in $R(\psi^{P_1})$, $R(\psi^{P_1}_{P_3})$ is orthogonal to the orthogonal complement of $R(\psi^{P_3})$ in $R(\psi^{P_2})$, $R(\psi^{P_2}_{P_3})$. In other words, the three subspaces are mutually orthogonal.

To prove the orthogonality of subspaces corresponding to signals of exactly period P the following theorems are introduced.

Theorem 1: For any two specific period P and Q ($P \neq Q$) and signal R of length L (L is a multiple of P and Q), let $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_m$ be all the possible divisors of P and Q respectively (including 1 and excluding P and Q, respectively). Then $R(\psi^P_{p_1\ldots p_n})$ and $R(\psi^Q_{q_1\ldots q_m})$ are orthogonal.

Proof: Without loss of generality, let $p_1 = q_1$ be the greatest common divisor of P and Q. Then, $R(\psi^P_{p_1\ldots p_n}) \subset R(\psi^P_{p_1})$ and $R(\psi^Q_{q_1\ldots q_m}) \subset R(\psi^Q_{q_1})$. By Lemma 1, $R(\psi^P_{p_1})$ is orthogonal to $R(\psi^Q_{q_1})$.

The next theorem proves that the subspace corresponding to signals of exactly period P is $R(\psi^P_{p_1\ldots p_n})$,

Theorem 2: Let $p_1, p_2, \ldots, p_m$ are the divisors of P (including 1 but not P). Then, S is exactly period P if and only if $S \in R(\psi^P_{p_1\ldots p_n})$.

Proof: First assume that $S \in R(\psi^P_{p_1\ldots p_n})$. For any period $\overline{P} < P$, $\overline{P}$ and P have same common divisor. Let p_1 be the greatest common divisor of $\overline{P}$ and P. From theorem 1, $R(\psi^P_{p_1})$ is orthogonal to $R(\psi^{\overline{P}}_{p_1})$. With $R(\psi^P_{p_1\ldots p_n}) \subset R(\psi^P_{p_1})$, $R(\psi^P_{p_1\ldots p_n})$ is orthogonal to $R(\psi^{\overline{P}})$.

Second, assume that S is exactly period P. By definition $S \in R(\psi^P)$ and $S \perp R(\psi^{\overline{P}})$ for any $\overline{P} < P$. In other words, S is in the orthogonal complement of $R[\psi^{p_1} \ldots \psi^{p_n}]$ inside $R(\psi^P)$. Therefore, $S \in R(\psi^P_{p_1\ldots p_n})$.

Calculation of orthogonal projection: To calculate the projection of signal R onto the orthogonal subspaces corresponding to exactly period P, let us define $p_1, p_2, \ldots, p_n$ be all the possible divisors of P (including a but not P). By definition, the subspace corresponding to signals of exactly period P is the orthogonal complement of the union of the subspaces corresponding to exactly period p_i inside $R(\psi^P)$. Then theorem 1 guarantees that

$$R(\psi^P) = \text{subspace of exactly period } P \oplus \sum_{\Theta} \text{subspace of exactly period } p_i \qquad (2)$$

3. TANDEM REPEAT DETECTION ALGORITHM

The main objectives of any tandem repeat identification algorithm are to identify its periodicity, its pattern structure and its copy number. A major difficulty in identifying tandem repeats arises due to the presence of inexact tandem repeats. An inexact repeat is one in which the substrings are similar, but not identical, e.g., ACGACGACC. Inexact repeats are thought to be representation of historical events associated with sequence. Thus, it is important for any tandem repeat identification technique to identify both exact and inexact tandem repeat structures in a DNA sequence.

A DNA sequence consists of a series of four nucleotide symbols A, C, G, T. However, when the nucleotides of a DNA sequence are mapped to some numeric value the tandem repeat identification problem becomes period detection problem. In [14] the EPSD technique was applied to identify the periods that were present in several synthetic and vibration data. In this paper we have successfully applied EPSD technique for identifying the tandem repeat structures in

Table 1. Nucleotide subsequences of a DNA sequence

	A	C	A	A	G	T	A	C	A	G	T	C	C	T	T
$S_A[n]$	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0
$S_C[n]$	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0
$S_G[n]$	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
$S_T[n]$	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1

DNA sequences. The complete input data was taken for identifying the periods in [14], however for tandem repeat identification problem we have applied a window based approach because the tandem repeat structure forms only a small portion of the given input DNA sequence. Sliding window EPSD technique helps in locating the correct position of tandem repeats in the DNA sequence.

In this section we present our tandem repeat identification algorithm. Our algorithm addresses the problems that were present in periodicity explorer algorithm (PE) [7]. The problems with PE algorithm are as follows:

- The nucleotide mapping is taken as follows: A=1+j1, C= –1+j1, G= –1– j1, and T=1–j1. Let the two DNA sequences be ACACACGT and ACACACTG. Using equation (14) from [7] the corresponding periodogram coefficients obtained for period-2 are 0.25 and 0.625 respectively. By comparing the two DNA sequences we observe that even though the two DNA sequences have equal degree of period-2 component, the periodogram coefficient obtained are different. This shows that the periodogram coefficient cannot act a good estimator for measuring periodicity. This problem has occurred due to arbitrary mapping of the nucleotides.
- Periodicity transform gives non-orthogonal decomposition of the signal. This result in different ways to decompose the signal which leads to different solution. As a result of this the PE algorithm is designed to be executed separately for every period. For example, if the PE is searching for repeats till say P_{max} (maximum period) then the algorithm has to be executed P_{max} times. This means that the run time of the PE algorithm is O(NWP_{max}) where N is the length of analyzed DNA sequence and W is the window size.
- PE algorithm cannot tell whether the tandem repeat present in the DNA sequence is of period P or multiple of P (i.e. 2P, 3P,… and so on).

We describe next, our proposed algorithm based on EPSD technique. Our algorithm takes care of the shortcomings in the PE algorithm. The algorithm is divided into three steps.

Step 1: Convert the DNA sequence $S[n]$ into four nucleotide subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$

The four nucleotide subsequences are obtained as follows:

$$S_\Omega[n] = \begin{cases} 1, & \text{if } S[n] = \Omega \quad \text{where } \Omega \in \{A,C,G,T\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

For example, Table 1 shows the $S_\Omega[n]$ components for the DNA sequence 'ACAAGTACAGTCCTT'.

Table 2. Calculation of repeat coefficient for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$

1. Accept window length (W), maximum period (P_{max})
2. **for** n=1 to N+W-1 **do** // N is the length of DNA sequence
3. $S_\Omega[n,\cdots,n+W-1] = S_\Omega[n,\ldots,n+W-1] - MEAN(S_\Omega[n], S_\Omega[n+1],\ldots S_\Omega[n+W-1])$
4. $\alpha_\Omega[1,\cdots,P_{max}] = EPSD(S_\Omega[n,\ldots n+W-1], P_{max})$
5. $\pi_\Omega[1,\cdots,P_{max}] = \dfrac{\alpha_\Omega[1,\cdots,P_{max}]}{\|S_\Omega[n,\ldots,n+W-1]\|^2}$
6. OUTPUT(j, $\pi_\Omega[j]$), where $\pi_\Omega[j] \leftarrow \max(\pi_\Omega[1],\ldots\pi_\Omega[P_{max}])$
7. $n = n$+1

od

Step 2: Calculation of tandem repeat coefficient for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$

Our algorithm is designed to identify tandem repeats from period-2 to maximum period (P_{max}) provided by the user within an observation window of length W. Like other signal processing techniques [5, 6, 7] our algorithm also use a sliding window based approach for identifying the position of the tandem repeats in DNA sequences. The algorithm for calculating period with maximum energy for the input DNA sequence of length N and input parameters (P_{max}, W) is provided in Table 2. At the end of this step we obtain a tuple <p, π > for each window where 'p' is the maximum period and 'π' is the repeat coefficient value.

Step 3: Identification of tandem repeat position and characterisation

In this step a threshold parameter (τ) is accepted from the user and then output of the subsequences obtained in step 2 are processed together to calculate repeats. The repeats obtained are represented by tuples $\langle \psi, i, l, p \rangle$ and the symbols in the tuples are defined as follows:

- $\psi \in$ power set of Ω, where $\Omega \in \{A,C,G,T\}$ & $\Omega \neq \Phi$
- i is the starting position of repeat
- l is the length of repeat.
- p is the period of repeat

The repeats $\langle \psi, i, l, p \rangle$ must satisfy the following conditions:

a. $\pi_\beta[i],\ldots,\pi_\beta[i+l-1] \geq \tau \quad \forall\ \beta \in \psi$

b. $p \bmod p_\beta[i] = 0 \quad \forall\ \beta \in \psi$ where $p \in \{p_\beta[i]\}$

The tuples $\langle \psi, i, l, p \rangle$ represents an exact tandem repeat *iff*

DNA Sequence = GAGTGCGTGCGTGC
G-G-G-G-G-G-G- ; -A----------; ---T---T---T--; -----C---C---C

Initial Repeats:	**Final Repeats**
<G,1,14,2>	<G,1,5,2>
<C,6,14,4>	<T,4,5,4>
<T,4,14,4>	<(G,C,T), 6,14,4> (Tandem Repeat)

Figure 1 – Repeat tuples calculation

$|\psi| \geq 2$ otherwise it represents some inexact-tandem type of repeats. One example is shown in figure 1.

4. EXPERIMENTAL RESULT

To demonstrate the capabilities of the tandem repeat identification algorithm experiments were performed on some actual DNA sequences available on the National Centre for Biotechnology Information (NCBI) website (http://ww.ncbi.nlm.nih.gov). 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.

EXAMPLE 1: The analysis of Homo sapiens collagen gene, accession number NM_001847 of length 6574 bp (base pairs) containing weak tandem repeat pattern is provided in this example. The tandem repeat coefficient obtained for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for window size (W) = 80 and maximum period (P_{max}) = 20 is shown in figure 2. From figure 2 we notice that subsequence $S_G[n]$ have significant repeat coefficient value from 50-4200 nucleotide position while the other subsequences $S_A[n]$, $S_C[n]$, $S_T[n]$ have repeat coefficient value at about 0.3. This shows presence of repetitive element in $S_G[n]$. Figure 3 shows the presence of period of length 9 when threshold value for repeat coefficient was taken as 0.7. The details of the repeat identified in the DNA sequence by our algorithm and SRF is provided in table 3. However, when PE algorithm was applied on this DNA sequence it gave tandem repeat of period 9 and multiple of 9 (i.e., 18, 27 and so on). 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 by EPSD technique. From table 3 we observe that repeat pattern of period-9 due to nucleotide 'G' is present in the input DNA sequence from 250 to 4200 bp. When the minimum score set to 30 several inexact period-9 like ours were reported by TRF 4.0.

EXAMPLE 2: The analysis of Homo sapiens, GeneBank Locus: HSVDJSAT of length 1985 bp is provided in this example. The tandem repeat coefficient obtained for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for window size (W) = 100 and maximum period (P_{max}) = 50 is shown in figure 4 (only the portion with significant repeat coefficient is shown). From the figure it is clear that all the subsequence have significant tandem repeat coefficient value from 1100-1430 bp. For a threshold value of 0.65, periods of length 19, 29, 37, 41, 43, 47, and 49 were reported for the subsequences and

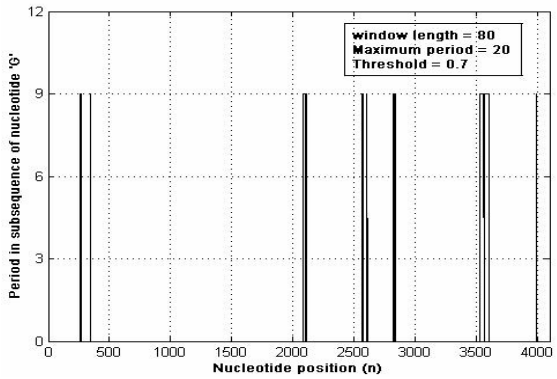

Figure 3 – Periods identified in the repeat coefficient sequence of $S_G[n]$ with window length=80, maximum period =20, and threshold = 0.7

Table 3 – Repeat pattern identified in H. Sapiens collagen gene

	Period	Starting position	Repeat length	Repeat pattern
Our Algorithm (window=80, max. period= 20, threshold = 0.7)	9	257	90	- g g - - - - - -
		347	82	- g g - - - - - -
		2084	105	- - - - - - - g g
		2565	123	g g - - - - - - -
		2824	97	- - g g - - - - -
		3528	158	- - - - - - g g -
		3988	81	- - - - - g g - -
TRF 4.0 (align. parameter= 2,5,7; min. score = 40, max. period = 500)	9	963	30	ggagaaaag
		1404	29	ccaggccca

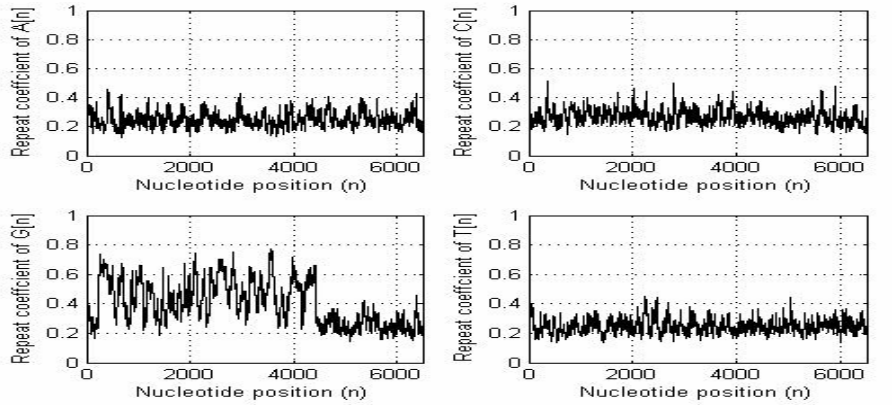

Figure 2 – Repeat coefficient of subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for window length = 80 and maximum period = 20

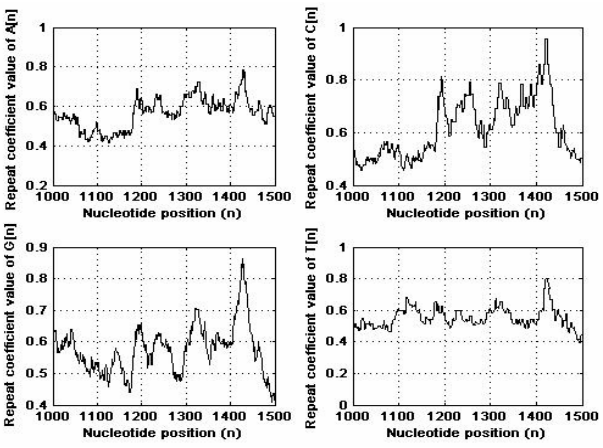

Figure 4 – Repeat coefficient of subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for window length = 100 and max. period = 50

tandem repeats of period length 19, 29 and 47 were reported. The tandem repeats regions were further processed and consensus pattern reported by our algorithm are aggctgggaggctggag, aggctgggatgctggag, tgggagaggctgggagag gctggga(t/g)(t/a)gc, gaggctgggagaggctgggagag-ctgggagaggct g-gattgctggga of period length 19, 19, 29 and 49 respectively. The algorithm [9] reported tandem repeats of period length 19, 38 and other periods there were combination of period 10 and 9 were reported. The sequence was also analyzed by TRF 4.0 with default parameters. The repeats of period length 2, 19, 49 and 10 were reported. Apart from tandem repeat many other periods were also reported by our algorithm in each subsequence.

5. CONCLUSION AND FUTURE WORK

A difference in design of a tandem repeat identification algorithm between signal processing approach and other techniques is that the signal processing techniques require a threshold parameter but other techniques require parameters like edit distance, hamming distance and several other parameters for identifying desired tandem repeats. Since the knowledge of mismatch parameters in terms of edit distance, hamming distance and other parameters are not known to user in prior, the signal processing approach offers a better choice for identifying tandem repeat in such situations.

The algorithm presented in this paper is based on exactly periodic subspace decomposition technique for identifying tandem repeat structures in DNA sequences. The algorithm runs in $O(NW\lg W)$ and is computationally faster than PE algorithm which runs in $O(NWP_{max})$ where N is the length of the analyzed DNA sequence, W is the window size and P_{max} is the maximum period to be identified. Our algorithm also resolves the problems that were present in PE algorithm. The algorithm is designed to analyze each nucleotide sequence separately and later on result of individual nucleotides are combine together to report tandem repeats in DNA sequences. An advantage of using our algorithm is that in addition to finding tandem repeat our algorithm also reports the various periods that are present in individual subsequences. This helps in identifying weak tandem repeat structure as well as individual nucleotide periods present in the DNA sequence. The analysis of individual nucleotide periods could be helpful in understanding their role in DNA sequence. In future we would like to apply our algorithm for interspersed repeat that are present in DNA sequences.

REFERENCES

[1] R. R. Sniden, "Biological implications of the DNA structures associated with disease-causing triplet repeats," *Human Genetics*, vol. 64, no. 2, pp. 346–353, 2000.

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

[3] Y. Nakarmura, M. Leppert, P. O' Connell, R. Wolf, T. Holm, M. Culver *et al.*, "Variable number of tandem repeat (VNTR) markers for human gene mapping," *Science*, vol. 235, pp. 1616–1622, 1987.

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

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

[6] T. T. Tran, V. A. Emanuele II, G. T. Zhou, "Techniques for detecting approximate tandem repeats," in *Proc. ICASSP 2004*, Montreal, Canada, May 17–21 2004, vol. 5, pp. 449–452.

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

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

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

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

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

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

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

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

[15] W. A. Sethares, T. W. Staley, "Periodicity transform," *IEEE Trans. Signal Processing*, vol. 47, no. 11, pp. 2953–2964, Nov. 1999.