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

IDENTIFYING INVERTED REPEAT STRUCTURE IN DNA SEQUENCES USING CORRELATION FRAMEWORK

Ravi Gupta, 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, ankumfec, ksd56fec}@iitr.ernet.in

ABSTRACT

The detection of inverted repeat structure is important in biology because it has been associated with large biological function. This paper presents a framework for identifying inverted repeat structure present in DNA sequence. Based on the correlation framework, the algorithm is divided into two stages. In the first stage the position and length of contiguous inverted repeats are identified based on the input parameters using correlation function. Later on in the second stage maximal inverted repeats are constructed by merging of continuous inverted repeats.

The advantage of the framework is that it can be successfully used for identifying both exact and inexact inverted repeats, returning maximal inverted repeat. Additionally, the framework does not need the user to specify parameters which require knowledge of system details. Experiments were performed on various chromosomes of Saccharomyces cerevisiae (baker's yeast) genome data available at NCBI website and some of the typical results are presented in this paper.

1. INTRODUCTION

Signal processing techniques offer a great promise in analyzing and deciphering genomic data. The genomic information is inherently discrete in nature because there are finite numbers of nucleotides in DNA alphabet. The standard approach of representing the genomic information by sequences of nucleotide symbols in the sequence of DNA and RNA molecule, by symbolic codons (triplets of nucleotides), or by symbolic sequences of amino acids in the corresponding polypeptide chains limits the methodology of handling the genomic information to mere pattern matching or statistical procedures. By properly mapping nucleotide symbols to some numeric sequences [1, 2], signal processing techniques provides a set of novel and useful tools for solving relevant problems of genomic data. In recent years signal processing is making its presence felt in the emerging universe of genome exploration.

For last few decades, the major thrust of DNA and protein analysis has been on string matching, either with goal of obtaining a precise solution (e.g., with dynamic programming) or more commonly a faster solution (e.g., with heuristic techniques). However, these heuristic methods do not work well on repetitive structures. In DNA, most repetitions occur as tandem or reverse complement repeats. Reverse complement structure is also called inverted repeat. Inverted repeats are widespread in both prokaryotic and eukaryotic genomes [3, 4, 5], and have been associated with a large number of possible functions. Promoters, viruses, and eukaryotes all contain inverted repeats. Origins of DNA replication from higher eukaryotes, such as monkey and human, are also enriched in inverted repeats. Inverted repeats have been implicated in the regulation of initiation of DNA replication in plasmids, bacteria, eukaryotic viruses and mammals [6]. More details regarding about the roles of inverted repeat in human disease is presented in [7].

Thus, it is important to detect the inverted repeat structure in a DNA sequence. A major difficulty in identification of repeats arises from the fact that the inverted repeats present in DNA sequence can be either exact or inexact, and are of unspecified length. The detection of exact inverted repeat is simple and can be achieved in linear time but the detecting an inexact inverted repeat has proven to be challenging task.

Many techniques have been developed to identify the inverted repast. Suffix tree technique [8] transforms the inverted repeat detection problem to finding longest common extension subsequence. It identifies exact or inexact repeat with fixed number of mismatches in linear time. However, the technique becomes both complex and inefficient for finding inexact inverted repeat without any prior knowledge of mismatches due to substitution or insertion/deletion of nucleotides. Another technique for detecting approximate inverted repeats in nucleotide sequences is inverted repeat finder (IRF) [9]. IRF is a statistically based heuristic algorithm and the approach is similar to BLAST algorithm. The program detects candidate inverted repeats by finding short, exact, reverse complement matches of 4-7 nucleotides (k-tuples) between non-overlapping fragments of a sequence. A "center" position is defined for each k-tuple match. Inverted repeat finder detects "clusters" of k-tuple matches having the same or nearly the same center and falling within a small interval of sequence. Candidate inverted repeats are confirmed (aligned and extended) or rejected by computing Smith-Waterman style similarity alignment. IRF run against a genome sequence using parameters match, mismatch, indel, and minimum score. Major drawback of this technique is the requirement of input parameters listed above. The result is technique is dependent on the input parameters. A user has to do many trial sessions specifying different set of values for these parameters in order to get a good match.

EINVERTED is another program available at *http://bioweb.pasteur.fr/seqanal/interfaces/einverted.html* is also used for finding inverted repeats in a DNA sequence. The algorithm is based on dynamic programming methodology. It works by finding alignments between the sequence and its reverse complement that exceeds a threshold score. Gaps and mismatches are assigned penalty (negative score). Matches are assigned a positive score. The score is calculated by summing the values of each match, the penalties of each mismatch and the large penalty of any gaps. Any region whose exceeds the threshold value are reported.

In this paper, we provide a correlation based signal processing technique in order to locate and identify the inverted repeat structures in DNA sequences. The inverted repeat detection algorithm is divided into two stages. It takes input parameter as the maximum length of the inverted repeat and minimum contiguous matches. These parameters, unlike in other search systems, do not require user to be expert in understanding the inner details of the system.

The paper is organized as follows. Section II describes about the correlation function and how it can be applied to detection of inverted repeat problem in DNA sequences. Section III presents an inverted repeat detection algorithm for DNA sequences. 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. CORRELATION FUNCTION

Correlation is a mathematical tool to quantify the degree of interdependence of one data upon another. In other words, it is used to establish the similarity between one set of data and another. The process of correlation occupies a significant place in signal processing. Applications of correlation are found in image processing for robotic vision or remote sensing by satellite in which data from different images is compared, in detection and identification of noise, and in many other fields, such as, climatology. The correlation for N-point data is given as

$$r_{12}[k] = \frac{1}{N}\sum_{n=1}^{N} f_1[n] f_2[n+k] \qquad (1)$$

where f_1 and f_2 are two functions for which the correlation is to be calculated. When $f_1[n] = f_2[n]$ then correlation is said to be auto-correlation and when $f_1[n] \neq f_2[n]$ then it is said to be cross-correlation.

The complexity of the correlation operation is O(N^2) which is quite expensive, especially when dealing with very large sequences. The correlation computation may be speeded up by exploiting the correlation theorem, usually stated as

$$r_{12}[k] = \frac{1}{N} F_D^{-1}(F_1[n] F_2^*[n]) \qquad (2)$$

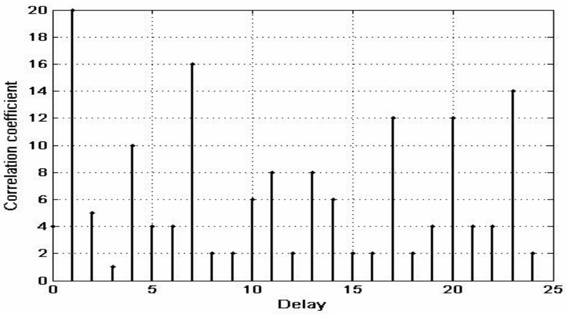


Figure 1 -The correlation coefficients obtained for sequences

where $F_1[n]$ and $F_2[n]$ are Fourier transforms (FT) of $f_1[n]$ and $f_2[n]$ respectively and F_D^{-1} is the Inverse Discrete Fourier transform (IDFT).

In this paper we provide a novel technique for identifying both exact and inexact inverted repeat structures in a DNA sequence. The technique is based on correlation of DNA sequence and its reverse complemented sequence. The value of correlation coefficients that are obtained after performing correlation is directly related to the presence of inverted repeat structures in the DNA sequence.

A primary step before performing correlation of DNA sequence is to assign some numeric values to DNA nucleotides. An arbitrary assignment would not give a correct correlation measure between the DNA sequences. For example, let A=1, C=2, G=3, T=4, and three sequences be $S1$ = AACC, $S2$ = AACG, $S3$ = AACT. By observation, the correlation between $S1$ and $S2$ is equal to the correlation between $S1$ and $S3$, since they are off by one element. However, using the given numeric mapping of the sequence, the correlation coefficient between $S1$ and $S2$ is 0.9, $S1$ and $S3$ is 0.82. In order to obtain a correct correlation measure we have converted the DNA sequence into four nucleotide subsequences. Similar nucleotide assignment has been used in [10, 11]. Consider a DNA sequence S of length L,

$$S = s_1 s_2 s_3 \cdots s_{L-1} s_L$$

consisting of L letters belonging to a given finite alphabet Ω. For DNA sequences, Ω = {Adenine (A), Cytosine (C), Guanine (G), and Thymine (T)}. The sequence S is divided into four nucleotide sequences S_A, S_C, S_G, and S_T. The sequences are calculated as follows:

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

In this way the original DNA sequence is decomposed into four binary sequences. The decomposition of the DNA sequence into nucleotide subsequences helps in obtaining correct correlation measure between the two DNA sequences. Consider two sequences S and S' as the given DNA and its reverse complemented sequences respectively. The correlation between the two sequences is calculated as follows:

$$r_{SS'}[k] = \sum_{\Omega} r_{S_\Omega S'_\Omega}[k] \qquad 0 \le k \le L-1 \qquad (4)$$

$$r_{S_\Omega S'_\Omega}[k] = \sum_{i=0}^{L-1} S_\Omega[i] \cdot S'_\Omega[i+k] \quad where \quad S'_\Omega[i] = S'_\Omega[i+L] \qquad (5)$$

The parameter $r_{SS'}$ gives the correlation between the DNA sequences and its reverse complemented sequence (i.e. S and S' respectively) and is used as a measure for inverted repeat sequence identification problem. The correlation coefficient obtained for a random DNA sequence AGCGGCATGATCA TGATCATGCCCG of length 25 is provided in figure 1. From the figure it is clear that there is high correlation for delay 1 and 7. The delay parameter in the correlation operation of sequences helps in identifying the position of the inverted repeat in the given DNA sequence and the value of correlation coefficient gives an idea about the number of matches in the inverted repeat.

3. INVERTED REREAT DETECTION ALGORITHM

The main objectives of any inverted repeat detection algorithm are to identify its length, its pattern structure and its position in the DNA sequence. The inverted repeat detection algorithm present in this section operates in two stages. The first stage identifies contiguous inverted repeats in the input DNA sequence. A contiguous inverted repeat is represented by tuple <*X*, *Y*, *l*> where (*X*, *Y*) represents a pair of coordinates revealing the position of the repeat in the genome sequence and *l* is the length of the repeat. In the second phase the small contiguous inverted repeats are merged to obtain inexact inverted repeats present in the DNA sequence.

Algorithm:
The inputs to the algorithm are a DNA sequence (*S*), minimum length of contiguous repeat ($L_{\min}$), window length (*W*).

Preprocessing Stage: As the signal processing techniques cannot operate on the symbolic data, hence an important task before applying a signal processing technique is to convert the symbols into number. In this step, four binary sequences (consisting of 0s and 1s) are constructed each for the input DNA sequence and its reverse complemented sequence as discussed in the previous section.

Identification of contiguous inverted repeat sequence: The major difficulties while detecting an inverted repeat in DNA sequence are the length of the repeat and the position of such repeat in the DNA sequence. As discussed in last section, the delay parameter in the correlation operation gives the location of inverted repeat and the value of correlation provide the number of matches, which can be used in finding the length of the inverted repeat. An exact repeat consist a single continuous repeat, however an inexact repeat consists of many small contiguous repeats. In this section the algorithm identify the location and length of small contiguous repeats present in the DNA sequence.

Table 1. Pseudo code for stage 1 of the algorithm

```
FIND INVERTED REPEAT ($S$, $Start$, $End$, $L_{\min}$ )
if CHECK FLAG($Start$, $End$) = TRUE
    then return
if ($End - Start + 1) < 2^* L_{\min}$
    then return
($N_A, N_C, N_G, N_T$)← number of nucleotide ‘A’, ‘C’, ‘G’, ‘T’ respectively in $S[Start\ldots End]$

$MaxMatch \leftarrow \min(N_A, N_T) + \min(N_C, N_G)$
SET FLAG ($Start$, $End$)
if $MaxMatch < L_{\min}$
    then return
$i \leftarrow Start,\ j \leftarrow End,\ TotalMatch \leftarrow 0$
while $i < j$ and $TotalMatch < MaxMatch$ and $S[i]=S[j]$
    do $i \leftarrow i+1, j \leftarrow j-1, TotalMatch \leftarrow TotalMatch+1$
if $TotalMatch >= L_{\min}$
  then OUTPUT ($Start$, $End$, $TotalMatch$)
        $Start \leftarrow Start + TotalMatch$
        $End \leftarrow End + TotalMatch$
        if $TotalMatch < MaxMatch$
          then FIND INVERTED REPEAT ($S$, $Start$, $End$, $L_{\min}$ )
        return
  else
        $Corr$ = FIND CORRELATION ( $S$, $I$, $Start$, $End$)
            // $I$ is the reverse complemented sequence of $S$
        $i \leftarrow 0$
        while $i < \lfloor WindowLength \rfloor$
            do if $Corr[i] >= 2^* L_{\min}$
                  then FIND INVERTED REPEAT ($S$, $Start$,
                                              $Start + i$, $L_{\min}$ )
                       FIND INVERTED REPEAT ($S$,
                                 $Start + i + 1$, $End$, $L_{\min}$ )
               $i \leftarrow i + 1$
        return
```

A pseudo code of this step is provided in the Table 1. The identification process is based on dividing the given DNA sequence into small subsequence until some stopping criteria is met. One of the important stopping criteria while searching for inverted repeats in a DNA sequence is based on the count of nucleotides A, C, G, and T in the sequence. The *MaxMatch* variable represents the maximum length of continuous inverted repeat sequence that can be present in the DNA sequence and is sum of min(N_A ,N_T) and min(N_C , N_G), where N_A , N_T , N_C , N_G are the counts of nucleotides A, C, G, T in the DNA sequence between *Start* and *End* position. For example, for AGCGGCATGATCATGATCATGCCCG, N_A=6, N_T =5, N_C =7, N_G=7 and *MaxMatch* =12 which mean at maximum there can a continuous inverted repeat of length 12 in the DNA sequence. So, for any DNA sequence if it is found that *MaxMatch* is less than $L_{\min}$ which is provided by the user can be straight away rejected and hence reducing our inverted repeat search cases. After all the stopping criteria for the DNA sequence fails, a search is made for an exact contiguous inverted repeat. If length of the length satisfies the minimum matching length criteria ($L_{\min}$) then its position and length is recorded otherwise a further search for inverted repeat is made in the DNA sequence.

The identification of the position and length of the inverted repeat is based on the value of the correlation coefficients that is obtained after performing a correlation between the DNA sequence and its reverse complemented DNA sequence. The delay parameter of the correlation tells the location of inverted repeat in the sequence and the value of correlation is directly related to the length of the inverted repeat.

Merging of contiguous inverted repeats: In this stage the contiguous inverted repeats that are present in the same window are merged together in order to form inexact inverted repeats. The output from previous stage consists a list of tuple <*X*, *Y*, *l*>. Two tuples <*X*1, *Y*1, *l*1> and <*X*2, *Y*2, *l*2> can be merged only if the following condition holds true:

$$X2 \geq X1 + l1 \quad \& \quad Y2 \leq Y1 - l1 \quad \textit{where } X1 < X2 \qquad (6)$$

For example, ACGGATATGT have contiguous inverted repeat as <1, 10, 2> i.e., AC – – – – – – GT and <5, 8, 2> i.e., ATAT, so both can be combined according to the above rule to obtain an inverted repeat as AC– –ATAT– – GT. An acyclic graph is constructed in order to obtain inexact inverted repeats from a list of contiguous inverted repeats generated in the previous stage. The nodes of the acyclic graph are labeled as <*X, Y, l*> where *X* represents starting location, *Y* the end location and *l* the number of matches. An edge is created from node *N*1≡<*X*1, *Y*1, *l*1> to node *N*2≡ <*X*2, *Y*2, *l*2> if and only if the condition provided in equation (6) holds true.

After the construction of graph is completed, a topological sorting [12] of the acyclic graph is done. The sorting may result in various paths and each such path forms an inverted repeat of the DNA sequence. For selecting the starting node for topological sorting the following conditions must be satisfied:

- starting node must be non-traversed node.
- if *P* ≡ <*X*1, *Y*1, *l*> is selected as a starting node then *X*1 must be the smallest starting location from the set of non-traversed node and *Y*1 must be farthest among all nodes that are starting from *X*1.

After reaching an end node, all the nodes of the current path are displayed in the order they were visited. Each such path obtained forms an inverted repeat of the input DNA sequence.

4. RESULT

To demonstrate the capabilities of the inverted repeat detection algorithm, experiments are performed on some actual DNA sequences available on public databases. We provide some of the results obtained when the algorithm was tested on chromosomes of *Saccharomyces cerevisiae* (baker's yeast) genome data. The contiguous repeats are shown in bold and are also underlined.

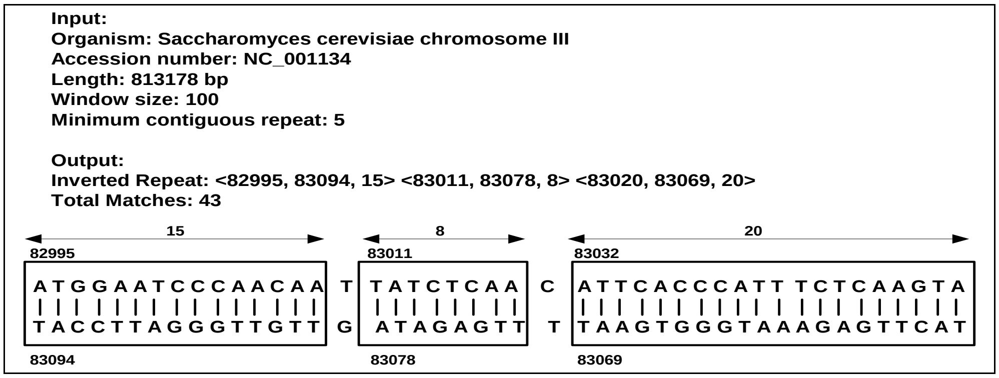
Saccharomyces cerevisiae chromosome III:

A detailed test was performed for different window sizes and minimum contiguous repeat length. The inverted repeat with highest number of matches reported when applied to chromosome III of *Saccharomyces cerevisiae* with window size equal to 100 and minimum continuous repeat length as 5 is shown in figure 2. Total number of matches in the reported inverted repeat was 43. When the window size was increased to 300 the length of the inverted repeat reported was increased and is given by:

<82899,83191,11> <82914,83176,24> <82939,83151,12> <82952,83138,5> <82958,83132,23> <82981,83108,29> <83011,83078,8> <83020,83069,20>

TATGTAGAAAT ATAG **ATTCCATTTTGAGGATT CCTATAT** C **CTCGAGGAGAAC** T **TCTAG** T **ATAT TCTGTATACCTAATATTAT** – **AGCCTTTATCAAT GGAATCCCAACAA** T **TATCTCAA** C **ATTCACC CATTTCTCAAGTA** CTATTCATCT **TACTTGAGAAA TGGGTGAAT** T **TTGAGATA** G **TTGTTGGGATTC CATTGTTGATAAAGGCT** A **ATAATATTAGGTAT ACAGAATAT** G **CTAGA** G **GTTCTCCTCGAG** C **ATATAGGAATCCTAAAATGGAAT** TAGC **ATTTC TACATA**

The starting location of the inverted repeat in the DNA sequence is 82899 and the length is equal to 293. The inverted repeat is obtained by merging eight contiguous inverted repeats of length 11, 24, 12, 5, 23, 29, 8 and 20. The contiguous repeats are shown in bold and are also underlined. The

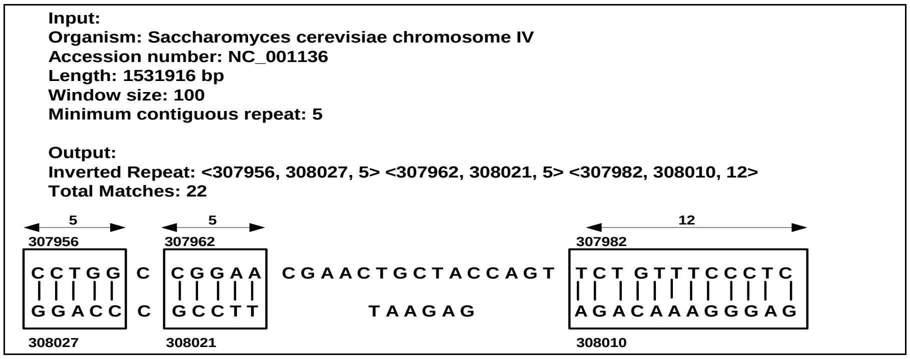


Input:
Organism: Saccharomyces cerevisiae chromosome III
Accession number: NC_001134
Length: 813178 bp
Window size: 100
Minimum contiguous repeat: 5

Output:
Inverted Repeat: <82995, 83094, 15> <83011, 83078, 8> <83020, 83069, 20>
Total Matches: 43

15
8
20
82995
83011
83032
A T G G A A T C C C A A C A A
T
T A T C T C A A
C
A T T C A C C C A T T T C T C A A G T A
T A C C T T A G G G T T G T T
G
A T A G A G T T
T
T A A G T G G G T A A A G A G T T C A T
83094
83078
83069

Figure 2 - An inverted repeat reported by the algorithm in chromosome III of Saccharomyces cerevisiae DNA sequence.



```
Input:
Organism: Saccharomyces cerevisiae chromosome IV
Accession number: NC_001136
Length: 1531916 bp
Window size: 100
Minimum contiguous repeat: 5

Output:
Inverted Repeat: <307956, 308027, 5> <307962, 308021, 5> <307982, 308010, 12>
Total Matches: 22
```

Figure 3 -An inverted repeat reported by the algorithm in chromosome IV of Saccharomyces cerevisiae DNA sequence.

total number of matches in the inverted repeat sequence is 132. This chromosome when tested by *EINVERTED* program gives out the same result when the window size was taken to be equal to that what we have taken.

Saccharomyces cerevisiae chromosome IV:

The inverted repeat with maximum number of matches reported by the algorithm for window size=100 and minimum contiguous repeat length = 5 is the following:

<307956, 308027, 5> <307962, 308021, 5> <307982, 308010, 12>

The inverted repeat is shown in figure 3. It is formed by merging 3 contiguous repeats of length 5, 5 and 12. The starting location of the inverted repeat is 307956. The total number of matches in the inverted repeat sequence is 22.

5. CONCLUSION

A simple and efficient technique for identifying inverted repeat structure in DNA sequences using correlation based framework is presented in this work. The advantage of using the algorithm is that it can efficiently detect both exact and inexact inverted repeat present in DNA sequence. Another advantage of this algorithm is that it generates all possible inverted repeats present in DNA sequence satisfying minimum continuous repeat matching criteria fixed by the user. The algorithm also reports the location and the length of the inverted repeat detected in the DNA sequence. The inverted repeat detection algorithm requires the specification of two well understood parameters: maximum length of inverted repeat and minimum number of continuous match.

A further work can be carried out in devising scoring method for faster merging of contiguous inverted repeats in order to obtain a maximal inverted repeat. Also the work can be extended for identifying other repetitive structures such as tandem and dispersed repeat present in the DNA sequence.

REFERENCES

[1] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279-303, 2002.

[2] W. Wang, Don H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Signal Processing*, vol. 50, no. 3, pp. 628-634, March 2002.

[3] H. Ogata, S. Audic, C. Abergel, P. E. Fournier, and J. M. Claverie, "Protein coding palindrome are a unique but recurrent feature in rickettsia," *Letter: Genomic Research,* vol. 12, no. 2, pp. 808-816, 2002.

[4] M. D. LeBlanc, G. Aspeslagh, N. P. Buggia, and B. D. Dyer, "An annotated catalog of inverted repeats of caenorhabditis elegans chromosomes III and X, with observations concerning odd/even biases and conserved motifs," *Genomic Research*, vol. 10, no. 9, pp. 1381-1392, 2000.

[5] H. Ogata, S. Audic, C. Abergel, P. E. Fournier, and J. M. Claverie, "Selfish DNA in protein-coding genes of rickettsia," *Science*, vol. 290, pp. 347-349, October 2000.

[6] C. E. Pearson, H. Zorbas, G. B. Price, M. Zanna Hadjopoulos, "Inverted repeats, stem loops, and cruciforms: significance for initiation of DNA replication," *J. Cell. Biochem*. vol. 63, no. 1, pp 1-22, 1996.

[7] J. J. Bissler, "DNA inverted repeats and human disease," *Frontiers of Bioscience*, pp. 408-418, March 1998.

[8] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge, U.K.: Cambridge Univ. Press, 1997.

[9] P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, G. Benson, "Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes," *Genome Research*, vol. 14, no 10A, pp. 1861-1869, 2004.

[10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattachrya, R. Ramaswamy, "Prediction of probable genes by fourier analysis of genomic sequences," *CABIOS*, vol. 13, no. 3, pp. 263-270, 1997.

[11] 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.

[12] T. T. H. Cormen, C. E. Leiserson and R. L. Rivest, *Introduction to Algorithms*, MIT Press, Cambridge, MA, USA, pp. 477-484, 6th Indian edition, 2001.