2004 IEEE International Workshop on Biomedical Circuits & Systems

DETECTION OF PALINDROMES IN DNA SEQUENCES USING PERIODICITY TRANSFORM

Ravi Gupta[1], *Ankush Mittal*[1], *Vipin Narang*[2], *Wing-Kin Sung*[2]

[1](rgcsedec, ankumfec) @iitr.ernet.in,
Department of Electronics & Computer Engineering,
Indian Institute of Technology Roorkee,
Roorkee, India 247667

[2](vipinnar, ksung) @comp.nus.edu.sg,
Department of Computer Science,
National University of Singapore,
3 Science Drive 2, Singapore 117543

ABSTRACT

The detection of palindrome sequences in DNA sequences is important in Biology. Palindrome is also called inverted repeat and is an important element in the human genome. This paper presents a novel technique of applying modified periodic transform and signal processing methods in the domain of palindrome detection. Palindrome has recently received a lot attention in the research community due to its significance in identifying several key areas in a genome in relation to gene amplification, gene regulation and diseases.

1. INTRODUCTION

A palindrome is a sequence of letters/words which reads the same in forward as well as backwards direction, for example 'madam'. Palindrome sequences are crucial for gene regulation. They are frequent as transcription factor binding sites, such as the zinc finger dimer Gal4. Besides regulating protein production through mRNA intermediates, perfect and imperfect inverted repeats have structural roles in tRNA, ribozymes, and some ribonuclear proteins.

Palindrome is also crucial for DNA replication. The single-stranded bacteriophage G4 contains inverted repeats ranging from 20 to 44 base pairs long that are required for proper function of the origin of replication. The protein DnaG binds to such an inverted repeat and this interaction is required for the initiation of replication [1]. Palindromes, whether perfect or imperfect, are imbedded in genomes and have critical functions. Yet, these structures pose a special impediment to DNA replication fidelity and are associated with several human disease related to genes. Short inverted repeats in the mammalian genome have a critical role in the initiation of gene amplification [2, 3]. Therefore a number of techniques, such as [4], have been developed for palindrome detection.

Signal processing offers a great promise in analyzing genomic data [5]. However, signal processing techniques have not had a greater impact in analyzing genomic data since signal processing techniques have traditionally dealt with numeric sequences rather than characters or symbols. With the emergence of Genomic Signal Processing [6], both traditional and modern signal processing techniques have started playing an important role in processing and analyzing genomic and proteomic data.

The algorithm presented in this paper uses periodicity transform [7] method to detect a palindrome in a given sequence. The algorithm is computationally linear with the length of the given sequence. The algorithm solves the palindrome detection problem transforming it into a detection of tandem repeat problem. The algorithm first rearranges the given sequence, and then a 2-period sequence closest to rearranged sequence is found using periodicity transform method. The 2-period sequence is found by projecting the rearranged sequence onto subspace P_2 . If the 2-period sequence and the rearranged sequence are found to be equal, then this means the original sequence is palindrome.

2. PERIODICITY TRANSFORM

The periodicity transform offers a method which helps in detecting periodicities in a given sequence. This transform decomposes finite-duration sequences into a sum of periodic sequences by projecting it onto a set of 'periodic subspaces'. The periodic subspaces are not orthogonal, leading to decomposition which is not unique.

2.1. Periodic Subspaces

A sequence of real numbers, $S(k)$, is said to be p-periodic if there is an integer p such that $S(k+p)=S(k)$ for all integers k. Let P_p be the set of all p-periodic sequences, and P be the set all periodic sequences. Consider a sequence $S \subset P$ containing N elements. This can be considered to be a single period of N elements, *i.e.*, $S \in P_N \subset P$, and the goal is to locate

smaller periodicities within S. In order to locate smaller periodicities within S, the sequences must be projected on subspaces P_p for $p < N$ [2]. When S is close to some periodic subspace P_p, then there exists a p-periodic sequence S_p which is closest to S. This S_p is said to be the ideal choice for decomposing S. For every period p and time shift s, define the sequence $\delta_p^s(j)$ for all integers j such that

$$\delta_p^s(j) = \begin{cases} 1, & \text{if } (j-s) \bmod p = 0 \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

The sequences δ_p^s for $s = 0,1,\ldots,p-1$ are called p-periodic basis vectors since they form a basis for P_p. For example, the three-periodic basis vectors span the 3-period subspace P_3 is shown below

j	...	-3	-2	-1	0	1	2	3	4	...
$\delta_3^0(j)$	...	1	0	0	1	0	0	1	0	...
$\delta_3^1(j)$	...	0	1	0	0	1	0	0	1	...
$\delta_3^2(j)$	...	0	0	0	1	0	0	1	0	...

An inner product can be defined on the periodic subspaces as a $P \times P \to \Re$ function given by

$$\langle x, y \rangle = \lim_{k\to\infty} \frac{1}{2k+1} \sum_{i=-k}^{k} x(i)y(i) \qquad (2)$$

for arbitrary elements x and y in P. If $x \in P_{p_1}$ and $y \in P_{p_2}$, then the product sequence $x(j)y(j) \in P_{p_1}P_{p_2}$ is p_1p_2 -periodic, and the inner product is given as

$$\langle x, y \rangle = \frac{1}{p_1p_2} \sum_{i=0}^{p_1p_2-1} x(i)y(i) \qquad (3)$$

The basic properties of inner product can be easily verified.

A sequence S is said to be orthogonal to the subspace P_p if $\langle S, S_p \rangle = 0$ for all $S_p \in P_p$. Any two subspaces are orthogonal if every periodic basis vector in a subspace is orthogonal to every vector in the other subspace. However, it is important to note that the periodic subspaces P_p are not orthogonal to each other.

2.2. Projection Onto Periodic Subspaces

Consider an arbitrary sequence $S \in P$. Then, a minimizing vector in P_p is defined as $S_p^* \in P_p$ such that

$$\|S - S_p^*\| \le \|S - S_p\| \qquad (4)$$

for all $S_p \in P_p$. Thus, S_p^* is the p-periodic sequence that is closest to the original sequence S. The projection theorem [7] shows how S_p^* can be characterized as an orthogonal projection of S onto P_p. Applying the projection theorem, $S_p^* \in P_p$ can be expressed as a linear combination of the periodic basis vectors δ_p^s (eq. 1) as

$$S_p^* = \alpha_0\delta_p^0 + \alpha_1\delta_p^1 + \cdots + \alpha_{p-1}\delta_p^{p-1} \qquad (5)$$

where the unique minimizing vector is the orthogonal projection of S on P_p. Hence, $S - S_p^*$ is orthogonal to each of the periodic basis vectors δ_p^s for $s = 0,1,\ldots,p-1$, *i.e.*,

$$\langle S - S_p^* \rangle = \langle S - \alpha_0\delta_p^0 - \alpha_1\delta_p^1 - \cdots - \alpha_{p-1}\delta_p^{p-1}, \delta_p^s \rangle = 0 \quad (6)$$

After simplification, the coefficients α_s are obtained as

$$\alpha_s = p\langle S, \delta_p^s \rangle \qquad (7)$$

From the definition of inner product in eq. (3), α_s are given by

$$\alpha_s = p\frac{1}{pN} \sum_{j=0}^{pN-1} S(j)\delta_p^s(j) \qquad (8)$$

Using the definition of δ_p^s from eq.(1), we obtain α_s as

$$\alpha_s = \frac{1}{N} \sum_{k=0}^{N-1} S(s+kp) \qquad (9)$$

Let $\pi(S, P_p)$ represent the projection of S onto P_p. Then

$$\pi(S, P_p) = \sum_{s=0}^{p-1} \alpha_s\delta_p^s \qquad (10)$$

where δ_p^s are the p-periodic basis vectors of P_p. Note that $\pi(S, P_p) \in P_p$. Also, note that when S is projected onto P_{np}, the best np-periodic component within S is determined. Therefore, the residual, $R = S - P_{np}$, has no np-periodic component. The projection onto P_{np} also captures the entire p-periodic component in R.

3. PALINDROME DETECTION ALGORITHM

The Palindrome detection algorithm is based on the idea that if we rearranged the second half of the sequence and if the rearranged sequence is 2-periodic then this shows that the given sequence is a palindrome. The algorithm begins with rearrangement of the sequence, and later checks whether or not the rearranged sequence is 2-periodic. Fig. 1 illustrates how a palindrome sequence is equivalent to a 2-periodic sequence after rearrangement.

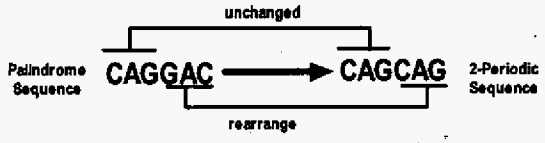

Fig. 1. Conversion of Palindrome sequence to 2-Periodic sequence

The algorithm is divided into three major sections. In the first section we rearrange subsequence of DNA sequence in a particular order. The rearrangement is done so that we can apply the periodicity transform method to the rearranged subsequence to find out the period of a particular length.

In the second section we obtain a periodic signal which is closest to the processed subsequence. We use the periodicity transform to find out a periodic signal of a particular period length.

In the final section we calculate Periodogram coefficient which is same as that calculated to find tandem repeats using short-time periodicity transform [8].

Algorithm:

Initial step: The given DNA sequence is divided into subsequences, where length of a subsequence is the length of palindrome which we have to detect. Let the given input DNA sequence be *I*, and let *S* be one of the subsequences of length *L* (length means the number of character present in *S*). The following steps are repeated for each such subsequence in *I* in order to detect whether *S* is Palindrome or not.

A. *Processing of subsequence*

The position of characters numbered from 1 to $\lfloor L/2 \rfloor$ in *S* is unchanged. But the position of characters in the sequence numbered from $L-\lfloor L/2 \rfloor+1$ to *L* in *S* is rearranged. The rearrangement is done by exchanging the position in the following manner

$$a_L \leftrightarrow a_{L-\lfloor L/2 \rfloor+1},\ a_{L-1} \leftrightarrow a_{L-\lfloor L/2 \rfloor+1+1},\cdots,$$

$$a_{L-\lfloor L/4 \rfloor} \leftrightarrow a_{L-\lfloor L/2 \rfloor+1+\lfloor L/4 \rfloor}$$

Let *U* denotes the rearranged sequence *S*.

If *U* is of odd length then remove the central character in *U* and reduce the length of *U* by 1.

B. *Calculation of 2-Periodic sequence*

The nucleotides of *U* are assigned a complex value. This very important because signal processing methods deal only with real or complex values. The mapping used in this paper is the following,

A→ 1+j, T→ 1 - j, C→ -1+j, G→ -1-j

Set period length, p = Half-Length *U*.

Use periodicity transform to calculate the 2-periodic sequence which is closest to *U*. Let it be called as *V*.

C. *Calculation of Periodogram Coefficient*

Calculate $\lambda = \dfrac{\| V \|^2}{\| U \|^2}$, where λ is the Periodogram coefficient and $\lambda \le 1$. If the value of λ is unity then sequence *V* is exactly 2-periodic, which further means that *S* is a perfect palindrome.

If we want to detect an imperfect palindrome, then we can set some threshold value for λ up to which the sequences can be accepted as a palindrome.

Time Complexity of the Algorithm:

The total number of subsequences of length *L* in a sequence of size *N* is $N-L+1$. The time complexity of finding the Periodogram coefficient is $O(L)$. Hence the total time complexity is given by $O((N-L+1)*L)$

4. EXPERIMENTAL RESULTS

The Palindrome Detection Algorithm was applied to two categories of DNA sequences. First, the algorithm is applied to a Pseudo-Random DNA Sequence generated by a Pseudo-Random generator program [9]. The second category is an actual DNA sequence.

A. *Pseudo-Random DNA Sequence*

The Pseudo-Random DNA Sequence of size 100 nucleotides was generated using Pseudo-Random generator program. The sequence is shown below. Table I gives the Palindromes and their positions found in the sequence. The results provided in Table I are only for a few palindrome sizes.

Sequence:
ACCACGCCGCTCACGTGGACCTAGATCCTCTCG
TGATGCATCATATGATGGCAACGATTTCGCGAC
AACCTAGGGTCCCTCGAAGTATTTTAATTTCAA

TABLE I
PALINDROMES DETECTED IN LENGTH 100 RANDOM DNA SEQUENCE

Length of Palindrome	Starting position of Palindrome in DNA sequence	Palindrome word found
3	4, 13	CAC
	6, 9, 62	CGC
	11, 29, 31, 80	CTC
	16, 34	GTG
	24	AGA
	30	TCT
	44	ATA
	45, 89	TAT
	59, 89, 90, 95	TTT
	63	GCG
	66	ACA
	74	GGG
	78	CCC
6	6	CGCCGC
	52	GCAACG
	88	ATTTTA
	91	TTAATT

B. Testing on Actual DNA Sequence

The algorithm was applied to several actual DNA sequences provided on the National Centre for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov). Result of one actual DNA is provided in Table II. The result provided here is only for few palindrome sizes.

TABLE II
SOME OF THE PALINDROMES DETECTED IN DNA SEQUENCE (accession number = L05934)

Accession number = L05934, Organism : Zea mays		
Length of Palindrome	Starting Position of Palindrome in DNA sequence	Palindrome word found
20	5960	CCTCCTCCTC CTCCTCCTCC
25	4774, 4776, 4778, 4780, 4782, 4784, 4786, 4788, 4790, 4792, 4794, 4796, 4798, 4800, 4802, 4804, 4806, 4708, 4710, 4712	ATATATATAT ATATATATAT ATATA
	4775, 4777, 4779, 4781, 4783, 4785, 4787, 4789, 4791, 4793, 4795, 4797, 4799, 4801, 4803, 4805, 4807, 4709, 4711, 4713	TATATATATA TATATATATA TATAT
	5974	TCCTCCCCCT CCCCCTCCCC CTCCT

5. CONCLUSION

The Palindrome detection algorithm can be used for detecting palindromes of any size present in DNA sequence. The algorithm can also be used to detect imperfect palindromes by setting some threshold value for Periodogram coefficient. Imperfect palindromes are those palindromes which need only few characters to be changed to make it a perfect palindrome. The experimental results show the effectiveness of our palindrome detection approach. No prior assumption of palindrome length need to be made and the algorithm is computationally fast.

6. REFERENCES

[1] J.J. Bissler, "DNA inverted repeats and human diseases," *Frontiers of Bioscience*, 3, pp. 408-418, March 1998.

[2] H.Tanaka, S.J.Tapscott, B.J.Trask, and M.C. Yao, "Short inverted repeats initiate gene amplification in mammalian cells," *Proc. Natl Acad Sci USA*, vol. 99, no. 13, pp. 8772-8777, June 2002.

[3] C.T.Lin, W.H.Lin, Y.L. Lyu and J.W. Peng, "Inverted repeats as genetic element for promoting DNA inverted duplication: implication in gene amplification," *Nucleic Acids Research*, vol. 29, no. 17, pp. 3529-3538, 2001.

[4] A. Apostolico, D. Breslauer and Z. Galil, "Parallel detection of all palindromes in a string," *Theoretical Computer Science*, vol. 141, no. 1-2, pp. 163-173, 1995.

[5] V. Veljkovic, I. Cosic, B. Dimitrijevic and D. Lalovic, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?", *IEEE Trans. Biomed. Engg.*, vol. 32, no. 5, pp. 337-341, 1985.

[6] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Mag.*, vol. 18, pp. 8-20, 2001.

[7] W. Wang and D.H. Johnson, "Symbolic signal processing," *IEEE Trans. Signal Processing*, vol. 47, pp. 2953-2964, November 1999.

[8] M. Buchner and S. Janjarasjitt, "Detection and Visualization of Tandem Repeats in DNA Sequences," *IEEE Trans. Signal Processing*, vol. 51, pp. 2280-2964, September 2003.

[9] W.H.Press, B.P.Flannery, W.T.Vetteriling, S.A.Teukoisky, *Numerical Receipes in C*, Cambridge University Press.