I J C T A, 9(17) 2016, pp. 8635-8642

Generalized idea of Palindrome Pattern Matching: Gapped Palindromes

Shreyi Mittal[*] and Sunita Yadav[**]

ABSTRACT

The process of palindrome works in a manner such that a string reads the same, when read in forward and backward directions respectively. Palindrome pattern matching is a wide field, finding applications in both theoretical and practical context. Applications include natural languages such as word combinatory and in bio-informatics. Gapped Palindrome or Quasi Palindromes are the special extensions of the palindrome, defined as one having the space between the arms of the given palindromic string, further categorized as the Length- constrained- palindromes and the Long-armed-palindromes. In this paper, we present the overall idea of the concept of the gapped palindromes, their extension as the Biological Gapped Palindromes along with the algorithms used to compute the same.

Keywords: Palindromes; Biological Gapped Palindromes; Long Armed Palindromes; Length Constrained Palindrome; Inverted Suffix Array; Dynamic Suffix Array; Enhanced Suffix Array.

I. OUTLINE

A Palindrome pattern matching is a traditional problem, with a varied lot of variations arising out of different practical scenarios. As known, palindrome is a string that reads the same forward and backward. String w is a palindrome if $w=yay^R$ where y is a string, y^R is a reversal of y, and 'a' is either a single character or the empty string. [1, 2, 3, 4] String length n is called palindromic rich, if it contains (n+1) distinct palindromes. A palindrome in a string is said to be maximal if it can't be extended in the inward or outward direction, while preserving a palindromic structure. [5,6] new paradigms are being designed for analyzing as well implementing the pattern matching study based on the palindromic structures. Palindromes in strings have widely been studied both in theoretical and practical contexts, such as in word combinatorics and in bioinformatics [1, 7]. In word combinatorics, researchers made studies on palindromes occurring in Fibonacci words [8]. One of the applications of the palindrome pattern matching is the natural languages, where palindromes have fascinated researchers since ancient times [9]. It has received considerable interest of computer science researchers, because of importance in identifying palindromic structures for different algorithmic problems [10, 11]. Tomohiro *et al.* proposed a linear time algorithm based on Knuth Morris Pratt (KMP) [12] to solve the palindrome pattern matching problem. The problem is such that it finds the maximal palindromic structure in a given text string that matches structure of a pattern in time O (n+p), where n is length of text and p is length of the pattern.

Palindromic structure are of different types one of the version being classified as a set of Gapped Palindromes , defined as the palindromes having spaces between the left and right arm of the palindromic string. Gapped Palindromes are responsible for chromosomal fragility and neurological disorder in humans [13]. A gapped palindrome is a sub-word of the form vuv^T for some u, v, where v^T is the reverse of v. Occurrences of v, v^T and u are called respectively left, right arm of the palindrome and spacer (gap) between them. Gapped palindromes have been categorized into the set of two classes as the Long-Armed-Palindromes and the Length-Constrained-Palindromes, and find their applications in the biological sequences. [14] In

[*] Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad - 201009, Uttar Pradesh, India; *E-mail: shreyi.calling22@gmail.com; yadav.sunita104@gmail.com*

length-constrained, lower and upper length bounds on the spacer length are specified in such a way that MinGap ≤ |u| ≤ MaxGap, and a lower bound on the arm length such that MinArm d" |v|, where MinGap, MaxGap, MinArm are pre-defined constants. With reference to biological sequences [15], the definition of gapped palindromes vary slightly where in the form v u v^T, v^T is the reverse complement of v.

In bioinformatics, palindromes have shown various functions [16, 17, 18, and 19]. The structures formed by palindromic subsequences play a role in regulation of gene activity or other processes in cells. For example, hairpin based palindrome structures are known to be present in the close vicinity of genes contributing to their normal functioning, or to diseases [7, 17, 20, and 21]. Biological palindromes include palindromes in DNA and RNA [21, 22]. In the molecular biology, analysis of DNA is key area of today's research. Words consisting of the palindromic structures are useful in DNA and RNA sequences, as they help to view the molecular capacity, i.e. the double- stranded stems. Thus, in a way helps to recognize the basic and important functioning of living beings at the perceived biological level. Concerning with the DNA or RNA sequences, are formed by the following base pairs as, Adenine (A), Thymine (T) [where A is complementary to T, in case of DNA], Uracil (U) [where A is complementary to U, in case of RNA], Cytosine (C) and Guanine (G) [where C is complementary to G].[5,17,18] Thus, they help to determine the parameters of gene activities along with their growth in the cellular structure. Those transformations occur as a result of the spacer, thus instead of searching only the palindromic structures in the genome sequences.

The DNA palindrome is a sequence of nucleotide base that reads the same as its reverse complement (where A is complementary to T and C is complementary to G). For example, 'CGTTAGC' is DNA palindrome such that 'GCAATCG' is complementary string (where G replaced by C and vice versa, A replaced by T and vice versa) of CGTTAGC. An RNA palindrome (having slightly different molecular structure to that of DNA) is a sequence of nucleotide base that reads same as its reverse complement (where G is complementary to C and U is complementary to A). For example, 'GCUUCG' is RNA palindrome such that 'CGAAGC' is complementary string (where C replaced by G, and A replaced by U) of GCUUCG. Furthermore, structures of palindromes considered in DNA and RNA are different [23, 24].

II. LITERATURE REVIEW

The literature review is covers the basics of the pattern matching along with their class of biological palindromes known as the Quasi palindromes or the biologically gapped palindromes respectively.

2.1. Palindrome Parmeterized Matching

Palindrome pattern matching is a wide field which has fascinated the natural language since ancient times. Palindrome pattern matching problem is to find all positions 'i'1≤ i ≤ n, in a reference sequence s [1….n] such that s[i….i+m-1] and given pattern p[1….m] are palindromic equivalent. Natural language includes the studies of the word complexities, and the periodicities of words overall. Palindrome is considered to be a symmetric string that appears to be the same backward and forward, when read from left to right or right to left. Palindromes have been defined vividly, depending on the words, letters, numerals, or the ones without spacing [1, 25].

Palindromes consist of varied concepts such as that of the palindrome richness [8]. A string is said to be palindrome rich, if the string consisting of the string length n, poses (n+1) distinct palindromes. Also, the close studies between the palindrome richness and the Burrows-Wheeler-Transforms have already been discovered. [27]

Palindromes have also been classified being on their finiteness and infiniteness, along with their palindromic complexity and their extensions to finite words have also been proposed. Another concept regarding the palindromic structure is that of the palindromic complexity, which is defined as the number

of palindromic substrings of a given length in the string. Also, the concept of the maximal palindromes was studied where the problems can be solved in the linear time. A palindrome in the string is maximal if it can't be extended outward while preserving a palindromic structure. Considering the word combinatories, many palindromic structure studies have been made, which include palindromic factors of Fibonacci words, Sturmian words and Ternary-square-free words.

New paradigm of pattern matching is based on the palindromes in strings, have been introduced, where the two strings of same length m are said to be palindrome equivalent iff the length of the maximal palindrome at every center in the strings is equal. [28] Thus, to find the same, the new concept of the palindromic suffix trees was used. The palindrome-suffix tree is a compacted trie, which represent the various suffixes. Each internal node consists of 2 children along with the label of two distinct out going edges. The tree is designed such that each leaf of the palindrome suffix tree is labeled uniquely with integers in a structured manner such that the path from the roots to leaf determines the longest suffix. [1]

Palindromic structures are thus been studied vividly fascinating researches in natural languages and word combinatories as well. Thus, overall periodicities of words have been studied or the study on the palindromic complexity has been made, be it the new properties of palindromes closely related to the parameterized pattern matching or be it the problem of inferring a string from a given set of palindromic structures. [28]

2.2. Biological Gapped Palindromes

Palindromic structures have also been played a vital role in the genome researches, i.e., find its applications in the field of bioinformatics. [3] The DNA and RNA sequences have relevance with the similar structure called the quasi palindrome. A quasi palindrome is a pair of reverse complementary repeats in a given string that are separated by a number of characters, i.e. spaces. Many studies have been conducted on the quasi palindrome and results show that they may control male germ-line gene expression. Thus recognition of such structure draws more attention towards the computational problem.

Considering the gapped palindromes, they are responsible to determine a word structure of the form vuv^t, where strings v and v^t are called arms and string u is called as the gap or the spacer. Computations of gapped palindromes in strings with length constrained can be done in linear time. A tool called Inverted Repeats Finder is designed to identify the approximate gapped palindromes in a string, by keeping the Levenshtein distance between the arms to be at most k (constant) and the length of the gap (spacer) to be fixed say q (constant) . Thus, this problem also solves the incremental string comparison disorder in humans. [13] These palindrome structures in humans show a major role in determining parameters of gene activities leading towards new developments in cells. These have been problem. Also, the results prove that the problem can be solved in the linear time. [3] The gapped palindromes are responsible for the chromosomal fragility and neurological further classified into the sets of two natural classes as the:

1. Long armed gapped palindromes
2. Length constrained gapped palindromes

The first class of the gapped palindromes, the long armed palindromes, verifies the condition %v%d" %u%, i.e. the length of the palindrome arm cannot be less than the length of the spacer. The second class is the length constrained palindromes, specified by the lower and upper length bounds on the spacer length such that MinGap d" %v%d" MaxGap and a lower bound on the arm length such that MinArm d" %u%, where MinGap, MaxGap, MinArm are the pre-defined constants.

Biological palindromes consist of the palindromes in DNA and RNA responsible for recognizing the major functioning of living beings at the biological level. Any given sequence in the string is made up over the four base pairs: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) in case of DNA. DNA

sequence is a palindrome of the sequence nucleotide bases that reads same as its reverse complement, where C is complementary to G; A is complementary to T, and vice versa. Any given sequence in the string is made up over the four base pairs: C,A,G,U, that is Cytosine, Adenine, Guanine and Uracil respectively in case of RNA. RNA sequence is a palindrome of the sequence nucleotide bases that reads same as its reverse complement, where C is complementary to G; A is complementary to U, and vice versa. [3, 7, 21]

The structures of DNA and RNA considered are different. In RNA secondary structures are considered along with the G-U pair. [23, 24] Many studies have been conducted on the gapped palindromes and their concept regarding the natural classes as the long armed and the length constrained. Kolpakov and Kucherov [20] have proposed a theoretical framework or algorithms for determining the two classes of gapped palindromes long armed and length constrained that runs in O(n + s) for a constant size alphabet, where S is the number of output palindromes and n is the size of the text. The algorithms designed make use of the Lempel- Ziv - Factorization method and the Suffix Array method. Henstra [16] described a method that can efficiently calculate the weight of the given RNA sequence, along with the size and number of the possible gapped palindromes in the input sequence. The techniques used for the palindrome matching are explained in the next section.

III. TECHNIQUES USED

Palindrome pattern matching is a technique, which consist of finding the same patterned string from the giving sequence, thus finding it's applications in varied field such as natural languages and bioinformatics. Considering the biologically gapped palindromes, varied techniques have been used to determine the two natural class of the gapped palindrome, being the length constrained and the long armed palindrome. The techniques used are explained as:

3.1. Suffix Trees

Suffix trees commonly known as the position trees are the indexing data structures that are used to store the suffixes of the given sequence (text) as their keys and the position in the given sequence as their values. Suffix trees thus, provides for the fast processing of the palindromic strings. Suffix trees are generally the component tire formed similar to that of the binary tree structure, which represents the length of the longest palindrome for all the suffixes. Each internal node consists of 2 children, consisting of the 2 distinct out going edges label starting with the non- negative integers. The tree is designed in a manner such that it consists of the exactly n leaves. Each leaf is assigned a labelled uniquely such that the path from the root to leaf determines the longest palindrome [1: p-q+1]$, where p is the length of the given string, q denotes the suffix and $ marks the end of the string. Further these are maintained by the use of active points. It processes the $(i\text{-}1)^{Th}$ character of the given string in the ascending order and generates the suffix tree respectively [1]. The main disadvantage of the suffix trees faced is the fact that they require the most space and secondly they have a poor memory reference locality. This problem can thus be overcome by the help of the array data structure called as the suffix arrays.

3.2. Suffix Arrays

Suffix arrays are defined as the repetition of index numbers that gives the starting position of the suffixes in a string in the alphabetical order. They have proved beneficial in solving problems like that of the data compression and the information retrieval on the biological sequences analysis and palindromes discovery. For any given string first the suffixes of the string are computed and then are arranged in the alphabetical order accordingly. On the basis of which the lcp (longest common prefix) tables are formed. Suffix arrays are further classified based on their types of traversal respectively as following:

a) Bottom-up: responsible for the retrieval of the complete suffix tree, used in the MGA (Multiple Genome Alignment) algorithm.

b) Top-Down: responsible for the retrieval of the subtree, mostly used for the exact pattern matching.

c) Traversal: this is the technique responsible for the retrieval of the suffix tree, by the help of the suffix links.

3.3. Longest Common Prefix

The LCP tables are used so as to further enhance the efficiency of the suffix arrays. A table generally consists of the length of the longest common prefix of the two subsequent strings in the given array. The table is prepared by the comparison of the subsequent suffixes in the suffix array in a manner such that when any two suffixes in the array have a longest common prefix of some set of characters say x, then the longest common prefix of all the ordered pairs of suffix between the two suffix is at least x characters long, that is they are compared character by character.

3.4. Inverted Suffix Arrays

The inverted suffix arrays are an extension to the suffix arrays, defined as the repetition of an index of the suffix array. When there exists more than one input string for the inverted suffix arrays, then the computation is held in the form of pairs, where the first value determines the string to which the suffix belongs, and the second value determines the start position of the suffix in the string respectively. Thus, are two- dimensional consisting of the string along with its start position in the string.

3.5. Dynamic Suffix Arrays

The main motive behind the use of the dynamic suffix arrays is to efficiently find the maximal gapped palindromes along with their natural classes, as the maximal length constrained palindromes and the maximal long armed palindromes respectively. As the name suggests are based on the dynamic construction of the new suffix arrays from the already existing suffix arrays being generated by the given string. The processing is simple to understand, the suffixes of the given sequence is stored by their index number respectively, hence consumes less space, thus provide advantage over the suffix arrays. They also tend to improve the efficiency by the insertion and deletion of a new character respectively. Results have shown that the use of the dynamic suffix arrays have improved the execution time by over 57.89%. [29, 30]

3.6. Enhanced Suffix Arrays

The enhanced suffix arrays are designed so as to improve the efficiency of the efficient retrieval for the maximal length constrained and long armed gapped palindromes. They help in overcoming the drawbacks of the suffix arrays and are defined as the construction of the suffix arrays along with the enhanced longest common prefix tables and their variants. [31]

IV. ALGORITHMS

The basic framework or the generalized theoretical algorithms were given by Kolpakov and Kucherov [20] for finding the maximal gapped palindromes for the length constrained and long armed palindromes. The algorithm makes use of the 2 techniques to compute the same that is the suffix arrays and the longest common prefix method. The algorithm designed consists of the two main steps:

a) Pre-Processing: this is the basic step where the given string is checked for the equivalence relation that is the unique index is assigned to each position under its equivalence class.

b) Computation: This is the main step for which the maximality and other constraints are satisfied, along with the comparisons made with the subsequent suffixes of the given string.

The other algorithms thereafter being implemented make use of the various techniques such as the inverted suffix arrays, dynamic suffix arrays and enhanced suffix arrays making it more efficient for the retrieval of the gapped palindromes. [1, 20, 29, 30, 31] the algorithms consist of the basic three steps procedure:

a) Reverse Complement: the first basic step is to find the reverse complement of the given DNA sequence or the given input string, and to find the equivalence relation

b) Concatenation: The suffixes computed are concatenated with the special symbol $, which denotes the end of the string.

c) Construct Suffix Array: the arrays are constructed using any of the feasible technique, along with the longest common prefix table, as required.

d) Constraint check: This is the important step where the constraints are check, covering the maximality constraint, spacer constraint and the arm's length constraint respectively.

Also, the other algorithm being designed so far to compute the palindromic weight for the given RNA sequence is proposed by Henstra[16] , which represents the structures of the large RNA sequences, identified and defined by the size and number of the gapped palindrome being input as the given sequence. It makes use of the suffix array and longest common sub-word method for the fast retrieval of the given sequence.

V. CONCLUSION

This paper presents the basic generalized idea of the palindrome pattern matching, with the main focus on the biologically gapped palindromes, along with their natural classes as the length constrained gapped palindrome, and the long- armed gapped palindromes. The algorithms discussed are based on the use of the techniques such as the suffix arrays, inverted suffix arrays, dynamic suffix arrays, enhanced suffix arrays and the longest common prefix methods. These techniques ensure the maximality constraint, the arm-length and spacer constraint and determines it to be gapped palindromes or not. Thus helps in the analysis of the DNA sequence, which makes analysis for the biological palindromes easy. These algorithms takes advantage of the various techniques and helps in the linear computation of the given sequence, by consuming less space, time, along with providing the ease of implementations. [29, 30]

Table 1 and Table 2, shows the comparative study of the existing algorithms for the Long-Armed-Gapped-Palindromes and the Length-Constrained-Gapped-Palindromes respectively.

Table 1
Comparison of Existing Algorithms for the : Long Armed Gapped Palindromes
[1, 20, 29, 30, 31]

Technique	*Total base pairs in dataset*	*Palindromic Weight*		*Length of Palindro-mic arm*	*Spacer Length*	*Long Armed Gapped Palindromes found in Sequence*	*Time Taken by proposed LAGP Algorithm (in ms)*
		Number of Palindromes	*Size of Palindrome*				
Inverted Suffix Array	230bp	1	10	4	2	ATTTAAAAAT	1.3186818
Dynamic Suffix Array	230 bp	1	10	4	2	ATTTAAAAAT	1.483516
Enhanced Suffix Array	230 bp	1	10	4	2	ATTTAAAAAT	1.153846

Table 2
Comparison of Existing Algorithms for the: Length Constrained Gapped Palindromes [1, 20, 29, 30, and 31]

Technique	*Total base pairs in dataset*	*Palindromic Number of Palind-romes*	*Weight Size of Palind-rome*	*Palindro-mic Arm Length*	*Gap Length*	*Length Constrained Gapped Palindrome found in sequence*	*Time taken by Existing Algorithm [1] (in ms)*	*Time by LCGP Algorithm (in ms)*
Inverted Suffix Array	230 bp	1	10	3	4	GCGTTTACGC	1.534862	1.208791
Dynamic Suffix Array	32bp	1	10	3	4	GCGTTTACGC	1.043956	0.439560
Enhanced Suffix Array	32bp	1	10	3	4	GCGTTTACGC	1.043956	0.254921

VI. FUTURE SCOPE

The algorithms presented can be executed and implemented over the large dataset and tested on the standard DNA with the large number of base pairs. The focus can be laid up on the RNA sequences or the protein synthesis to determine the gapped palindromes along with their classes. In addition to the existing, the work can be extended for the biological sequences having the palindromes with the properties of insertion, deletions and substitutions as well.

REFERENCES

[1] Tomohiro I, Inenaga S, Takeda M. Palindrome Pattern Matching. Theoretical Computer Science 2012: 1-9.

[2] Groult R, Prieur E, Richomme G. Counting distinct palindromes in a word in linear time. Information Processing Letters 2010; 110 (20): 908- 912.

[3] Hsu PH, Chen KY, Chao KM. Finding all approximate gapped palindromes. Proceedings of the ISAAC 2009, vol. 5878 of LNCS: 1084–1093.

[4] Kolpakov R, Kucherov G. Finding Maximal Repetitions in a Word in Linear Time. Symposium on Foundations of Computer Science, IEEE Computer Society 1999: 596–604.

[5] Tomohiro, I, Shunsuke, I, Masayuki, T.: Palindrome Pattern Matching.

[6] Morris, J.H., Pratt, V.R.: A linear pattern-matching algorithm. Tech. Rep. 40, University of California, Berkeley (1970)

[7] Gusfield D. Algorithms on Strings, Trees, and Sequences.Cambridge University Press 1997, New York.

[8] Droubay X. Palindromes in the Fibonacci word. Information Processing Letters 1995; 55 (4): 217–221.

[9] Porto AHL, Barbosa VC. Finding approximate palindromes in strings. Pattern Recognition 2002; 35: 2581–2591.

[10] Apostolico A, Breslauer D, Galil Z. Parallel detection of all palindromes in a string.11th Annual Symposium on Theoretical Aspects of Computer Science vol. 775 of LNCS, Caen, France: Springer 1994: 497-506.

[11] Kolpakov R, Kucherov G. Finding repeats with fixed gap. Proceedings of the 7th International Symposium on String Processing and Information Retrieval, SPIRE, Acoruña: IEEE 2000: 162 - 168.

[12] Tomohiro I, Inenaga S, Takeda M. Palindrome Pattern Matching. Theoretical Computer Science 2012: 1-9.

[13] Gerald R. Smith. Meeting DNA palindromes head –to –head. Genes and Development, 2008; 22: 2612-2620.

[14] Kolpakov, R, Kucherov, G.: Searching for gapped palindromes. Theoretical Computer Science 410(51), 5365-5373(2009)

[15] Tevatia S, Prasad R. Multi-patterns Parameterized Matching with Application to Computational Biology. International Journal of Information and Communication Technology (Inderscience) 2014: In press.

[16] Henstra SJ. Determining gapped palindrome density in RNA using suffix arrays. Leiden Institute of Advanced Computer Science, Leiden University 2010: 1-15.

[17] Höhl M, Kurtz S, Ohlebusch E. Efficient Multiple Genome Alignment. Bioinformatics 2002; 18 Suppl. 1: S312–S320.

[18] Gerald R. Smith. Meeting DNA palindromes head –to –head. Genes and Development, 2008; 22: 2612-2620.

[19] Liu, Bin, Junjie Chen, and Xiaolong Wang. Application of Learning to Rank to protein remote homology detection. Bioinformatics, 2015: btv413.

[20] Kolpakov R, Kucherov G. Searching for gapped palindromes. Theoretical Computer Science 2009; 410 (51): 5365–5373.

[21] Gupta R, Mittal A, Gupta S. An efficient algorithm to detect palindromes in DNA sequences using periodicity transform. Signal Processing 2006; 86: 2067–2073.

[22] Porto AHL, Barbosa VC. Finding approximate palindromes in strings. Pattern Recognition 2002; 35: 2581–2591.

[23] Liu, Bin, et al. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics, 2015;31.8:1307-1309.

[24] Liu, Bin, et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic acids research,2015:gkv458.

[25] Groult, R., Prieur, E., Richomme, G.: Counting distinct palindromes in a word in linear time. Information Processing Letters 110(20), 908-912 (2010).

[26] Glen, A., Justin, J., Widmer, S., Zamboni, L.Q.: Palindromic richness. European Journal of Combinatorics 30(2), 510-531 (2009).

[27] Restivo, A., Rosone, G.: Burrows-Wheeler trasnform and palindromic richness. Theoritical Computer Science 410(30-32),3018-3026(2009)

[28] I. T., Inenaga, S., Bannai, H., Takeda, M.: Counting and verifying maximal palindromes. In: Proc. SPIRE2010. LNCS, vol. 6393, pp. 135-146 (2010).

[29] Salson M, Lecroq T, Leonard M, Mouchard L. Dynamic extended suffix arrays. Journal of Discrete Algorithms. 2010; 8: 241-57.

[30] S Gupta, S Yadav, R Prasad. Searching Gapped Palindromes in DNA Sequences using Dynamic Suffix Array. Indian Journal of Science and Technology, Vol 8(23),70645, 2015.

[31] Abouelhoda MI, Kurtz S, Ohlebusch E. The Enhanced Suffix Array and its Applications to Genome Analysis. Proceedings of the second Workshop on Algorithms in Bioinformatics, vol. 2452 of LNCS: Springer- erlag 2002: 449-463.