Mining and Analysis of Tandem Repeated Patterns in Oncogenic Sequences involved in Cancer progression

Satish Kumar*, Dharminder Kumar*, Ashok Chaudhury**
*Deptt. of Computer Science & Engineering
** Centre for Bio & Nano Sciences
Guru Jambheshwar University of Science & Technology, Hisar (Haryana)

Abstract- Tandem repeat patterns are very useful for biologists as they describes a pattern that helps to determine an individual's inherited traits. Tandem repeats can be very useful in determining parentage. These repeated nucleotides play a very important role to analyse and understand the various disorders available in cancer disease. Various data mining techniques like clustering, association analysis and classification etc. can be used for analysis of these repeated nucleotides.

Keyword –Tandem repeats Patterns, BWtrs, Bio-PHP

Introduction

The biological sequences consist of four nucleotides bases adenine (A), guanine (G), cytosine (C) and thymidine (T). These forms the complete DNA sequence of an organism. Many times these are repeated in a definite order forming a track of repeated units. Each of these units can range from 1-60 nucleotides. These repeated can be divided into two main types micro-satellite and mini-satellites. When 10 to 60 nucleotides are repeated , then the repeats are called mini-satellites and the repeats with fewer nucleotides are called mini-satellite. When exactly two nucleotides are repeated, it is called a dinucleotide repeat (for example: AGAGAG or AG_3). When three nucleotides are repeated, it is called a trinucleotide repeat (for example: CAGCAGCAGCAG or CAG_4), and abnormalities in such regions can give rise to trinucleotide repeat disorders and when the number is not known or variable, it is refer to as variable number tandem repeat (VNTR)[1,2].

These repeats are of very much importance as they help in determining the parentage of child's in legal cases, individual inheritance trait can be determined through them and they helps in developing the primers for the sequencing and amplification of biological characters[3]. These repeats also responsible for the particular functions of the proteins codes by the genes having the codon repetition of particular amino-acid, such as the case of DNA binding proteins. A lot of human disorder or diseases are also associated with these repetitive elements diseases such as Huntington's disease [4] and certain forms of Fragile X syndrome [5]. The change in the frequency of particular repeats can result in the development of disease such as cancer[6]. As the genes consist of both coding and non coding regions. The changes in the repeats in coding region are of more importance as that directly affects the organism in the form of abnormal proteins. Also there are many evolutionary constraints on protein coding regions by protein function compared with the constraints of non-coding regions. These constraints involves stabilization of the protein core structure. Out of different repeats tri-nucleotide repeats in the exons forms a identical run of amino acids. For example there are four codons coding for the amino-acid alanine: $(gca)_n$, $(gcc)_n$, $(gcg)_n$, and $(gct)_n$ This homogeneous tract of alanine destabilized the protein secondary structure[7].

During the disease development repeat stretches may also be subjected to elongation and shortening processes. Human genome consist of approximately 2% of the nucleotide sequences in the form of tandem repeats in which the length of the repeat unit is between 1 and 11 bp(8). The functional role of tandem repeats is poorly understood. They are, however, known to be involved in several genetic diseases and they can be successfully used as the genetic markers.

There are many software tools available for finding the tandem repeats which includes

1) **Phobos - a tandem repeat search tool for complete genomes**

Phobos tandem repeat finder was developed by Dr. Christoph Mayer in 2006. It is a tandem repeat search tool for complete genomes. PHOBOS can search for tandem repeats with a unit size of more than 5000 bp, which in the STAMP modules implies that primers can also be designed for minisatellites and tandem repeats with even longer units. Search settings and the output format of PHOBOS can be adjusted in a flexible manner, making it an ideal multipurpose tandem repeat search tool

2) **BWtrs: A tool for searching for tandem repeats in DNA sequences based on the Burrows–Wheeler transform-**

In this algorithm a new and very efficient web-based application for large scale exact searches for all tandem repeats in genomic sequences, based on the idea of backward search with the Burrows–Wheeler Transform (BWT) algorithm was presented. The Burrows–Wheeler Transform is an efficient data compression algorithm, which recently gains an increasing interest in the aspect of applications in genomics. This algorithm allows for listing all occurrences of exact tandem repeats in a given string of length n in O(n log n) time. It uses efficient string indexing structure by Ferragina and Manzini for searching for the occurrences of so called rearmost tandem repeats that are then used to list the locations of the desire preferred tandem repeats, namely, the maximal tandem repeats of the primitive motif BWTrs.

3) **Bio-PHP Method–**

This project was started in December 2005 by Dr. Joseba Bikaandi, a lecturer in the Department of Immunology, Microbiology and Parasitology, Faculty of Pharmacy, The University of the Basque Country,Vitoria-Gasteiz, Spain Joseba has developed several PHP scripts for simulation of molecular biology techniques, and a website has been developed with those scripts at insilico.ehu.es. The aim of Dr. Joseba is to make available most of the source code running at his site throw biophp.org. The aim of this site is to share knowledge by using a Wiki-like service.

In this paper, a new algorithm is proposed to determine the tandem repeat pattern in the genomic sequences. The algorithm works on exhaustive search to determine the sequence repeats of all size and frequency. For the study we have taken the protein coding sequences of fifty four genes involved in carcinogenesis.

Material and Method

An algorithm is developed which searches the tandem repeats in genomic sequences. The algorithm works in exhaustive manner which is time consuming but is very effective in finding the exact repeats of any length and frequency. Currently it is limited to find the repeat consisting monomer's of three to ten nucleotides. The algorithm work exhaustively without any approximation, this presents a true picture of internal repeats present in the sequence. A number of algorithms are there which work on some approximations and till date there is no program which works exhaustively and report all the available repeats. Algorithm is also effective in determining the nested repeats which span the common region.

To check the performance of algorithm, a set of fifty-six genes involved in cancer is taken as input.. These genes are hyper-methylated in the in carcinogenesis and maintains the cancer progression. The genes were extracted from the Cervical Cancer Gene Database(CCDB), which is a comprehensive collection of the genes involved in the progression and maintenance of cervical cancer(table 1). The protein coding region of these genes were downloaded from the from the NCBI nucleotide database in fasta formatted files. These coding regions were then searched for the internal repeats one by one using the algorithm GUI which is implemented in java. The repeated sequences results in formation of amino-acid tract in which a particular single or more amino acids are repeated a number of times. This repetition of the amino acids results in the formation of special secondary structures which can perform a special function of DNA binding or can also disrupt the core structure of the molecule. The algorithm output consist of the repeating unit(di-nucleotide or tri-nucleotide) , repeat frequency or copy number (how many times repeat occur in sequence) and start and end position of the repeat, positions are separated by a comma. If a single repeat is present at many locations in the same genomic sequence it reports all the positions as tab delimited. Figure 1 shows the graphical user interface and the results.

IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org

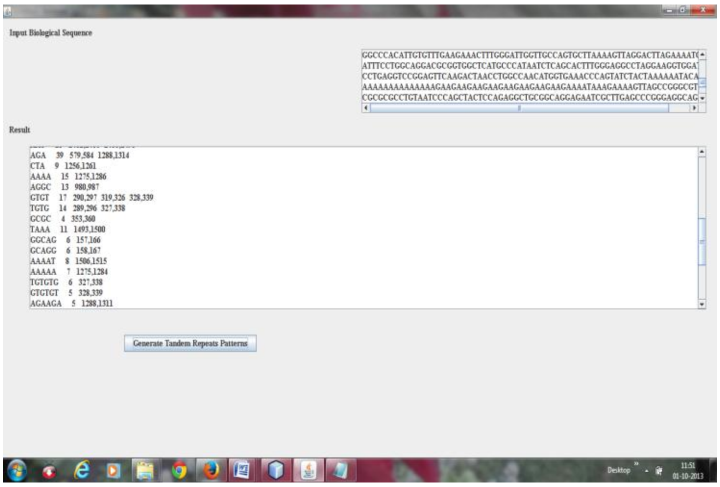


Fig 1 Showing the java graphical interface for the algorithm with the formatted output.

Result and Discussion

The protein coding region of the fifty four genes (GENE NAME TABLE)involved in the cancer progression downloaded from NCBI and were used to detect the tandem repeat pattern in the protein coding region (Table 2). The biological sequences consist of different kind of tandem repeats ranging from mono-nucleotide to repeats in which many nucleotides are repeated at adjacent positions. A variation in the repeat number or the frequency can result in the disruption of the protein function. As a large track of similar amino-acids may disrupt the core structure.

In the present study we developed an algorithm for finding the exact tandem repeats in the genomic sequences which is implemented in JAVA programming. The algorithm performs an exhaustive search for finding the exact tandem repeats. There are a number of other tandem repeat finders available which works on heuristic approaches which are BWTrs, Phobos and Bio-PHP etc. The algorithm works by forming a initial suffix matrix which is then used for identifying a tandem repeat in the sequence. The algorithm is trained in such a manner that it will report the repeats of any length and copy number given the sequence of any length. Because of its exhaustive approach we have presently limited our search to short tandem repeats which ranges from two to ten nucleotides.

In the 54 coding sequences the algorithm searches a total of 4741 short repeats, which is quit larger than the repeats identified by other similar tandem repeat finders. The repeats ranges from three nucleotides to eight nucleotides with a copy number ranging from 2-10. Repeats of higher order are less frequent in the coding regions. Table 1 shows the positional occurrence of different repeats in the input sequences, so the tri-nucleotide repeat occurs at 3300 times at different positions and their graphical representation is shown in Fig.1.

S. No	Repeat length	Occurrence
1	3	3300
2	4	843
3	5	279
4	6	212
5	7	78
6	8	29
	Total	4741

Table 1: Shows the occurrence of different repeats in fifty four sequences.

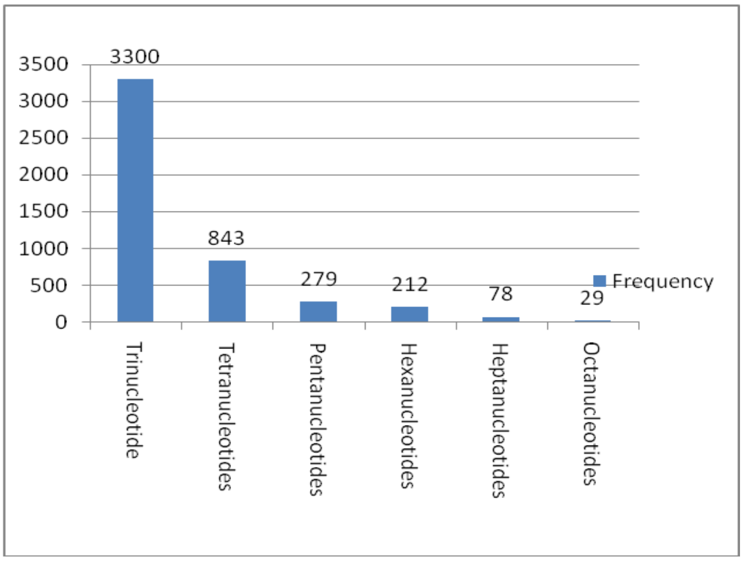


Fig.1- Graphical representation occurrence of different repeats in fifty four sequences.

Out of the total 4741 repeats there were 483 unique repeats whose distribution is given in Table 2. The study shows that the repeat of three nucleotides are more frequent with a count of 3300 and have a higher range of copy number which is upto 10 in the surveyed sequences. The dimeric repeats of three nucleotides occurs of 3112 times and there are 11 repeats with a frequency of ten nucleotides. After these four nucleotides repeats were more common with a count ranging upto 844. It consist of 793 dimeric repeats while repeats of higher order were less frequent with only two repeats of nine-nucleotide(table 3).

S.No.	REPEAT LENGTH	Frequency
1	3	64
2	4	162
3	5	136
4	6	92
5	7	18
6	8	11
	Total	483

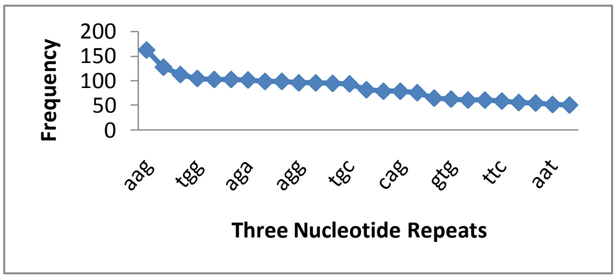
Table 2 showing the distribution of 483 unique repeats

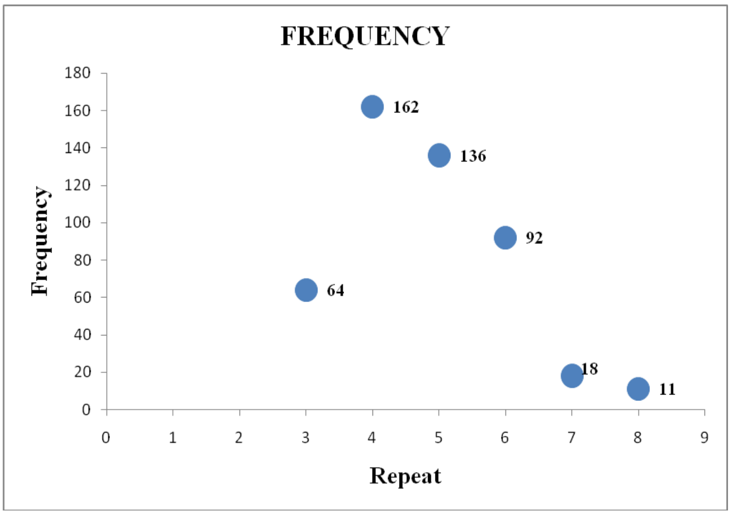
Repeats of Three Nucleotides	COUNT	Repeats of Four Nucleotides	COUNT
10	11	-	-
9	1	9	2
6	11	8	2
5	9	-	-
4	19	4	12
3	137	3	34
2	3112	2	793

Table 3 showing the distribution of 3 and 4 nucleotide repeats.

The three nucleotide repeats does not results in any change of amino acids so their copy number does not causes any reasonable change in sequence property while repeat of higher order can results in frameshift of amino acids. That why these repeats are less common in the coding region and occurs with more frequency in the inter-genic or non-coding region.

IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org





REFERENCES

1. Oki, E., Oda, S., Maehara, Y., Sugimachi, K. (1999). "Mutated gene-specific phenotypes of dinucleotide repeat instability in human colorectal carcinoma cell lines deficient in DNA mismatch repair". Oncogene 18 (12): 2143–2147.

2. Pennisi, E. (Dec 2004). "Genetics. A ruff theory of evolution: gene stutters drive dog shape". Science 306 (5705): 2172–2120.

3. Manasatienkij C, Ra-ngabpai C.Clinical application of forensic DNA analysis: a literature review.J Med Assoc Thai. 2012 Oct;95(10):1357-63.

4. Marcy E. MacDonald1, Christine M. Ambrose1, Mabel P. Duyao1, Richard H. Myers2, Carol Lin, Lakshmi Srinidhi1, Glenn Barnes1, Sherryl A. Taylor1, Marianne James1, Nicolet Groot1, Heather MacFarlane1, Barbara Jenkins1, Mary Anne Anderson1, Nancy S. Wexler3, James F. Gusella. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993 Mar 26;72(6):971-83.

5. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, Reiner O, Richards S, Victoria MF, Zhang FP, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome.Cell. 1991 May 31;65(5):905-14.

6. Xia X, Rui R, Quan S, Zhong R, Zou L, Lou J, Lu X, Ke J, Zhang T, Zhang Y, Liu L, Yan J, Miao X. MNS16A Tandem Repeats Minisatellite of Human Telomerase Gene and Cancer Risk: A Meta-Analysis. PLoS One. 2013 Aug 22;8(8):e73367

7. Borstnik B, Pumpernik D. Tandem repeats in protein coding regions of primate genes. Genome Res. 2002 Jun;12(6):909-15.

8. Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature.2001;409:860–921