Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Analysis of Tandem Repeat Patterns in Viral Oncogene (BRAF) Across Homologous Species

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

Abstract- Local repetitions in genomes are called tandem repeats. A tandem repeat contains multiple, but slightly different copies of a repeated unit. Tandem repeat patterns are very useful for biologists. In this paper, ten homologus species sequences are taken as input and various tandem repeats of upto nine nucleotides are extracted. These repeated nucleotides play a very important role to analyse and understand the various disorders available in various diseases. Various data mining techniques like clustering, association analysis and classification etc. can be used for analysis of these repeated nucleotides.

Keywords: **Tandem repeats, Suffix Matrix, BWtrs, Nucleotides**

I. Introduction

In nucleotide sequences two or more adjacent and approximate copies of a sequence of nucleotides are Tandem repeats. They are relatively common and, in primates, up to 10% can exist within protein coding genes (Jurka and Pethiyagoda, 1995). Different types of repetitive elements can account for up to 50% of the genome (Lander et al., 2001). The presence of tandem repeats and variations within these repeats have been associated with a large number of diseases and phenotypic outcomes but they also often exhibit population variability that is not manifestly detrimental in any specific allelic form. It is this variability that makes them useful markers in linkage analysis and DNA fingerprinting (Edwards et al., 1992, Weber and May,1989).

DNA tandem repeats within coding regions may or may not be observed as protein tandem repeats. For instance, coding CAG repeats implicated in repeat expansion diseases often translate as an expanded polyglutamine tract in the protein, but not all repeats translate in this way because of the redundancy of the genetic code; repeats that are approximate at the DNA sequence level might not appear as protein tandem repeats. This is an important observation because mixed codon repeats are more stable in evolutionary terms because they are less prone to slippage (Hancock and Simon, 2005). This is supported by the observation that variable tandem repeats are more likely to be homogenous, that is not to contain mismatches between adjacent copies of the tandem repeat in the repeat array (Wren et al.,2000).

Tandem repeats often have different names depending on their lengths; tandem repeats of unit length 1 to 5 nucleotides are generally termed microsatellites (Tautz and Schlotterer, 1994), whereas the term minisatellite generally describes repeats with unit lengths above this (Vergnaud and Denoeud, 2000). These definitions are somewhat arbitrary, though there is general consensus in the literature. But what exactly are tandem repeats, what is their purpose, and how might they have arisen? It has been discussed whether certain tandem repeats might serve as a mechanism for rapid adaptive response to various selection pressures, for instance in host defence genes (Wren et al., 2000). However, no strong evidence currently exists to support this hypothesis. It has been shown in prokaryotes that changes in repeats correlate with phenotypic changes that facilitate responses to various environments (Sylvestre et al., 2003) and it has been strongly suggested that repeats are an integral part of pathogen adaptation their hosts (Jordan et al., 2003, van Belkum et al., 1998, Verstrepen et al., 2005). Repeats may therefore be instrumental in facilitating rapid adaptation to new environments and thus act as evolutionary tuning knobs" (Li et al., 2004), a suggestion that is consistent with recent evidence that dramatic morphological changes in species of dog correlates with changes in repeats involved in developmental processes (Fondon and Garner, 2004), these changes presumably arising as a result of selection pressures from selective breeding. Other interesting roles have been postulated for repeats; complementary repeats have the potential to generate complementary hairpin loops. As secondary structure is known to constitute an important part of alternative splicing, and examination of the relationship between these repeats and alternative splicing found that 84% of complementary repeats were contained in the 44% of genes assessed that were known to have alternative splices (Lian and Garner,2005). Clearly, any variations within repeats essential in alternative splicing could potentially have dramatic effects on gene expression.

While SNPs are useful as markers and in disease association studies, repeats, by way of their innate mutability and evolutionary mechanisms, have the potential to introduce a greater amount of change at a DNA sequence level. When these changes arise in regions of functional importance, they are more likely to lead to significant changes in these regions. For example, one extra copy of a trinucleotide tandem repeat within a protein coding region can introduce an

extra amino acid into the protein sequence. However, one extra copy of a dinucleotide repeat can lead to a frameshift error (because protein sequences are translated as codons (3 bases) of the RNA transcript) and this can often cause disease (e.g. (Nobukuni et al., 1991)). It is reasonable to assume that the potential for repeats to generate significant changes in functionally important regions would render them more likely to be under functional constraints and indeed this is often the case. The various roles of repeats can broadly be classified into 3 categories (Li et al., 2002):

(a) Chromatin organisation, where they have roles in forming DNA and chromosomal structures and the organisation of centromers and telomeres.

(b) Regulation of DNA metabolic processes. Recombination can arise due to the DNAaltering effects of repeats. Mismatch repair system genes contain variable mononucleotide repeats in their coding regions, suggesting a role in the modulation of evolutionary mutation rate. Repeats can act as arrest sites during DNA replication and repeats can also affect cell cycle progression.

(c) Regulation of the activity of genes. Repeats can affect gene activity when located in promoters. They may also have significant roles in gene transcription given that genes related to transcription and signal transduction have an over-representation of repeats whereas genes for structural proteins do not. In terms of direct functional importance, tandem repeats have a number of important functions: repeats have important roles in stimulating transcriptional activity (Kashi et al., 1997).

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumour metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable.

II. Material and Method

A suffix matrix based 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.

In this paper, a database was created and compare the various tandem repeat patterns obtained by analysing the various viral oncogene across homogenous species. To check the performance of algorithm, a set of ten sequences of viral oncogenic is taken as input. These sequencers are given as:

1. Homo sapiens v-raf murine sarcoma viral oncogene homolog B (BRAF), mRNA
2. PREDICTED: Pan troglodytes v-raf murine sarcoma viral oncogene homolog B1, transcript variant 4 (BRAF), mRNA
3. PREDICTED: Canis lupus familiaris v-raf murine sarcoma viral oncogene homolog B1 (BRAF), mRNA
4. PREDICTED: Bos taurus v-raf murine sarcoma viral oncogene homolog B1 (BRAF), mRNA
5. PREDICTED: Rattus norvegicus v-raf murine sarcoma viral oncogene homolog B1 (Braf), mRNA
6. Gallus gallus v-raf murine sarcoma viral oncogene homolog B1 (BRAF), mRNA
7. Danio rerio v-raf murine sarcoma viral oncogene homolog B1 (braf), mRNA
8. Drosophila melanogaster pole hole (phl), transcript variant A, mRNA
9. Anopheles gambiae str. PEST AGAP004699-PA (AgaP_AGAP004699) mRNA, complete cds
10. Caenorhabditis elegans abnormal cell LINeage family member (lin-45) (lin-45) mRNA, complete cds

These coding regions were then searched for the internal repeats one by one using the algorithm GUI which is implemented in java netbeans. 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.

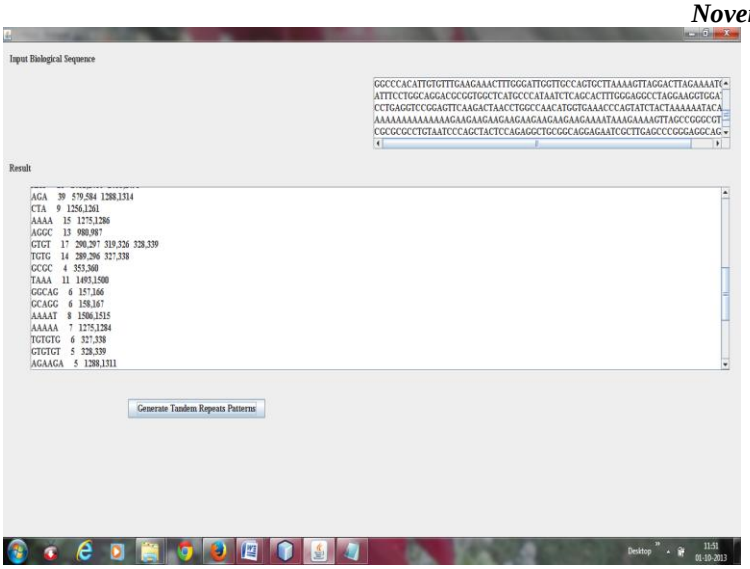

Fig 1 Showing the java graphical interface for the algorithm with the formatted output.

III. Result and Discussion

An suffix matrix based algorithm is developed 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 all the ten biological sequences the algorithm searches a total of 2154 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.

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

S. No	Repeat length	Occurrence
1	3	1604
2	4	249
3	5	117
4	6	176
5	7	2
6	8	6
	Total	2154

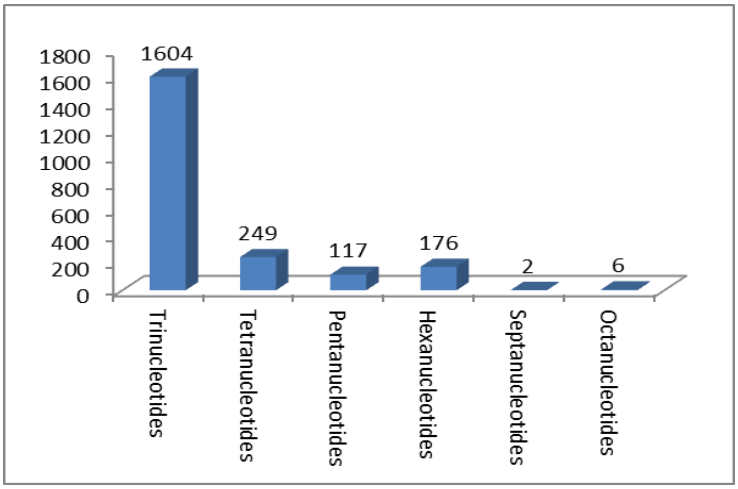

Fig.2- Graphical representation occurrence of different repeats in ten biological sequences.

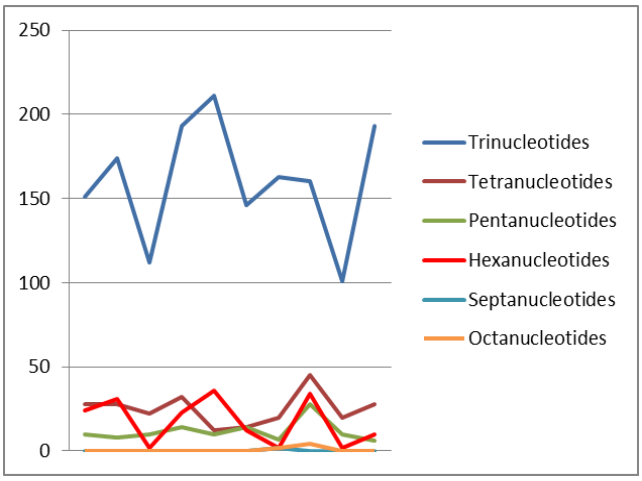

Fig 3- chart showing the trends of reapeated patterns over time

The study shows that out of the total 2154 repeats there are 232 unique repeats whose distribution is given in Table 2. The study shows that the repeat of four nucleotides are more frequent with a count of 76 and have a higher range of copy number which is upto 10 in the surveyed sequences. The dimeric repeats of four nucleotides occurs of 249 times and there are 6 repeats with a frequency of ten nucleotides. After these three nucleotides repeats were more common with a count ranging upto 70. It consist of 1604 dimeric repeats.

S.No.	Repeated Nucleotides	Frequency
1	3	70
2	4	76
3	5	29
4	6	53
5	7	1
6	8	3
	Total	232

Table 2 showing the distribution of 232 unique repeats

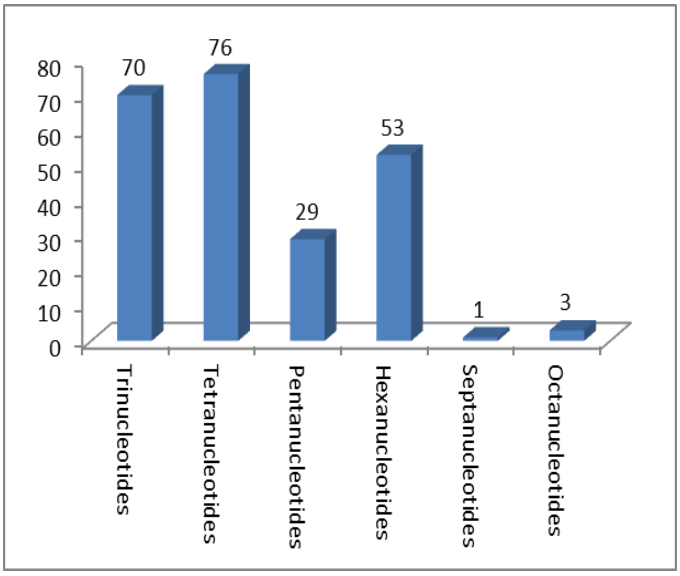

Fig.4 – Graphical representation of unique repeats.

In this analysis it was found that the higher nucleotide repeats are less common in Homologous species. The trinucleotides and pentanucleotides are more common and repeats of septa and octanucleotides are less frequent as shown in fig. 4. 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.

REFERENCES-

1. Jurka, J. & Pethiyagoda, C. (1995) Simple Repetitive Dna Sequences From Primates: Compilation And Analysis. *J Mol Evol,* 40**,** 120-6.
2. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S.,Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., Mcpherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C.,Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E.,Frazier, M., Et Al. (2001) Initial Sequencing And Analysis Of The Human Genome. *Nature,* 409**,** 860-921.
3. Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T. & Chakraborty, R. (1992) Genetic Variation At Five Trimeric And Tetrameric Tandem Repeat Loci In Four Human Population Groups. *Genomics,* 12**,** 241-53.
4. Weber, J. L. (1990) Informativeness Of Human (Dc-Da)N.(Dg-Dt)N Polymorphisms. *Genomics,*7**,** 524-30.
5. Hancock, J. M. & Simon, M. (2005) Simple Sequence Repeats In Proteins And Their Significance For Network Evolution. *Gene,* 345**,** 113-8.H
6. Jordan, P., Snyder, L. A. & Saunders, N. J. (2003) Diversity In Coding Tandem Repeats In Related Neisseria Spp. *Bmc Microbiol,* 3**,** 23.
7. Li, W. (1997) The Study Of Correlation Structures Of Dna Sequences: A Critical Review. *Computchem,* 21**,** 257-71.
8. Taneda, A. (2004) Adplot: Detection And Visualization Of Repetitive Patterns In Complete Genomes. *Bioinformatics,* 20**,** 701-8.
9. Verstrepen, K. J., Jansen, A., Lewitter, F. & Fink, G. R. (2005) Intragenic Tandem Repeats Generate Functional Variability. *Nat Genet,* 37**,** 986-990.
10. Willingham, A. T. & Gingeras, T. R. (2006) Tuf Love For "Junk" Dna. *Cell,* 125**,** 1215-20.
11. Ki, 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.
12. Ennisi, E. (Dec 2004). "Genetics. A Ruff Theory Of Evolution: Gene Stutters Drive Dog Shape". Science 306 (5705): 2172–2120.
13. Manasatienkij C, Ra-Ngabpai C.Clinical Application Of Forensic Dna Analysis: A Literature Review.J Med Assoc Thai. 2012 Oct;95(10):1357-63.
14. A. Merkel, N. Gemmell, Detecting Short Tandem Repeats From Genome Data: Opening The Software Black Box, Brief. Bioinform. 9 (5) (2008) 355–366.
15. D. Ames, N. Murphy, T. Helentjaris, N. Sun, V. Chandler, Comparative Analyses Of Human Single- And Multilocus Tandem Repeats, Genetics 179 (2008) 1693–1704.

