Available online at www.jbr-pub.org

Open Access at PubMed Central

JBR

The Journal of Biomedical Research, 2014, 28(5):416-422

Research Paper

RSMD-repeat searcher and motif detector

Udayakumar Mani✉, Vaidhyanathan Mahaganapathy, Sadhana Ravisankar, Sai Mukund Ramakrishnan

Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA University, Thanjavur, Tamil Nadu 613401, India.

Received 01 May 2013, Revised 27 July 2013, Accepted 30 August 2013, Epub 20 March 2014

Abstract

The functionality of a gene or a protein depends on codon repeats occurring in it. As a consequence of their vitality in protein function and apparent involvement in causing diseases, an interest in these repeats has developed in recent years. The analysis of genomic and proteomic sequences to identify such repeats requires some algorithmic support from informatics level. Here, we proposed an offline stand-alone toolkit Repeat Searcher and Motif Detector (RSMD), which uncovers and employs few novel approaches in identification of sequence repeats and motifs to understand their functionality in sequence level and their disease causing tendency. The tool offers various features such as identifying motifs, repeats and identification of disease causing repeats. RSMD was designed to provide an easily understandable graphical user interface (GUI), for the tool will be predominantly accessed by biologists and various researchers in all platforms of life science. GUI was developed using the scripting language Perl and its graphical module PerlTK. RSMD covers algorithmic foundations of computational biology by combining theory with practice.

Keywords: motif, repeats, genomic sequence, proteomic sequence, computational biology, combination algorithm

INTRODUCTION

With the avaialbility of complete genome sequence of many organisms and significant reduction in sequencing costs, the volume of biological data has been increasing exponentially. Bioinformatics emerging as a multi-disciplinary field has aided in the organization and assembly of these biological data in a more comprehensive manner. Identification of motifs and sequence repeats which are integral parts of DNA and protein sequences has been improved using bioinformatics techniques. Sequence motif is a nucleotide or amino acid sequence pattern that is widely observed across genomic data displaying strong biological significance. Identifying a motif helps in the formation of special secondary structures which may provide structural mechanism[1-3]. Contradictory to its essentiality, sequence repeats are also seen to be associated with a growing number of neurological disorders and diseases[4,5]. Studies on trinucleotide repeats showed that 9 neurological disorders are caused by an increased number of CAG repeats when they are present within the coding regions of genes[6]. This stresses the need to identify and analyze various repeats occurring in sequences.

Repeats can be classified into highly repetitive sequences and moderately repetitive sequences. The moderately repetitive sequences can be further classified into tandem repeats and interspersed repeats. Microsatellites are common among all sub-classes of tandem repeats. They are associated with various disease genes and have been used as molecular markers in linkage analysis and DNA fingerprinting studies, also seemingly playing an important role in genome evolu-

✉Corresponding author: Udayakumar Mani, Department of Bioinformatics, School of Chemical and Biotechnology, Shanmugha Arts Science Technology & Research Academy, SASTRA UNIVERSITY, Thanjavur, Tamil Nadu 613401, India. Tel/Fax: +04362-264101 Ext. 189/+04362 264120, E-mail: uthay@bioinfo.sastra.edu.

doi: 10.7555/JBR.28.20130065

tion[7]. Identification of microsatellites, inter simple sequence repeats (ISSRs) and directed amplification of minisatellite DNA (DAMD-PCR)[8–10] are rarely used among the acclaimed DNA-marker technologies that process vital information with repeats as the pinnacle point. Microsatellite repeats mostly vary from 3 to 6 nucleotides/amino acids. Trinucleotide repeats are causative for various neurological disorders and are also an important classification in microsatellites.

PROSITE[11], Repeat Finder[12], TAMO[13] and STAR[14] are few existing online softwares and servers which aid in the identification of motifs and repeats in proteins and DNA. With the increase of trinucleotide neurological diseases, identification of such repeats has become of great importance. Despite the availability of many tools, softwares and servers for the identification of repeats and motifs, there is not any tool for the identification of such disease causing trinucleotide repeats. Softwares and tools are required to incorporate a user-friendly environment, which allows researchers to easily understand, extract and navigate the available data. These are major drawbacks that were encountered while users accessed the above mentioned tools and softwares.

For the detection of microsatellites and trinucleotides, we have developed a user-friendly offline tool-repeat searcher and motif detector (RSMD). RSMD not only facilitates the identification of repeats but also aids in identification of sequence motifs.

CONSTRUCTION OF RSMD

As the relative ease Perl helps biologists to understand the language, it has been extensively employed along with its graphical module PerlTK in developing RSMD. The tool is divided into 2 main frames, and respectively sub-divided into 4 row frames and 2 graphical user interface (GUI) frames. The first row frame provides access to the major modules of the tool. The second row frame acts as the input panel. The third row frame contains legends for the pictorial representation of the sequences given as input. It also contains a feature to introduce mutations in the uploaded sequence. The final row frame displays the graphical representation of the input sequences. The first GUI frame displays various repeats along with its frequency of occurrence in the input sequence. Corresponding positions of the repeats are graphically represented in the second GUI frame (***Fig. 1***).

Repeat Searcher, Motif Detector, Similarity Search, Prosite Pattern Finder are the four main modules of RSMD. The tool contains additions features such as local and global sequence alignment (***Table 1***).

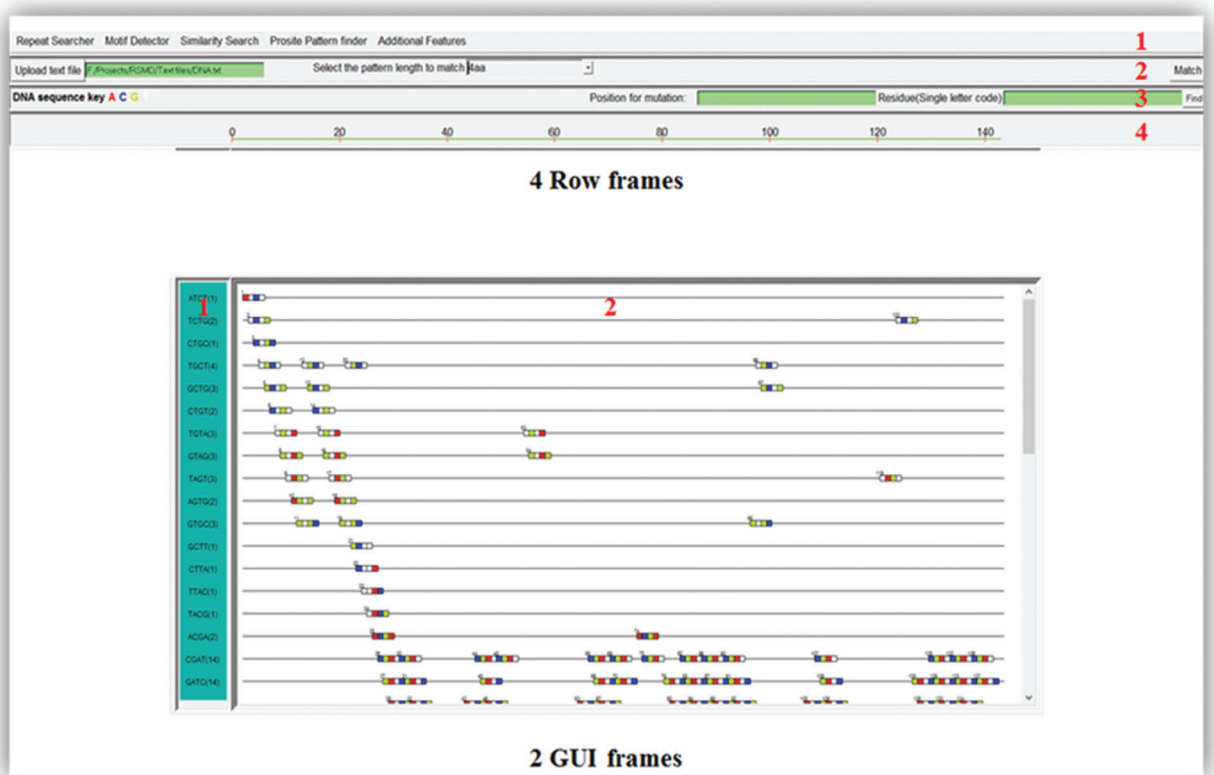


Fig. 1 **Frames of RSMD.** 4 row frames and 2 GUI frames. RSMD, Repeat Searcher and Motif Detector.

Table 1 **The usage of various modules in RSMD.**

No	Module name	Use
1	Repeat searcher	Identification of repeats in single/multiple sequence (.fasta)
2	Motif detector	Identification of motifs in a single sequence (.fasta)
3	Similarity search	Identification of repeats from an ClustalW alignment file (.aln)
4	Prosite pattern hunter	Identification of motifs from a single sequence using prosite patterns (.fasta)

WORKING OF RSMD

Maneuvering RSMD starts by selecting the required module from available options. On selecting the required module, users will be requested to upload the file which contains the input DNA or protein sequence (s) (.fasta or .aln file). On uploading the input file, users must select a suitable window length (1, 2, 3, 4, 5 or 6 amino acids/nucleotides). The entire analysis of input sequence is performed based on this window length. The sequence is searched for unique repeats, and its occurrence in the sequence is graphically depicted in GUI frames. The analysis of the sequence is performed and the presence of any motifs lying within is also depicted. All the motifs are obtained from PROSITE and other literature resources[15,16]. The motifs are already stored in a database that enables easy and efficient use of toolkit. Disease-causing motifs can also be inferred from the uploaded sequence. Mutation module bases can be altered at various positions and the altered graphical map can be visualized. Each nucleotide or amino acid is represented with a unique color code, thus making the result easier to interpret. The detailed workflow of RSMD is shown in ***Fig. 2***.

ALGORITHM OF RSMD

RSMD makes use of a disambigous combination algorithm that unifies it from its existing counter parts. For the input sequence, the algorithm works on 3 main parameters, including number of sequences, window length and sequence length. On uploading the input sequence(s), the algorithm separates all the sequences into separate array indices. The sequence with the maximum length is taken and made as the reference.

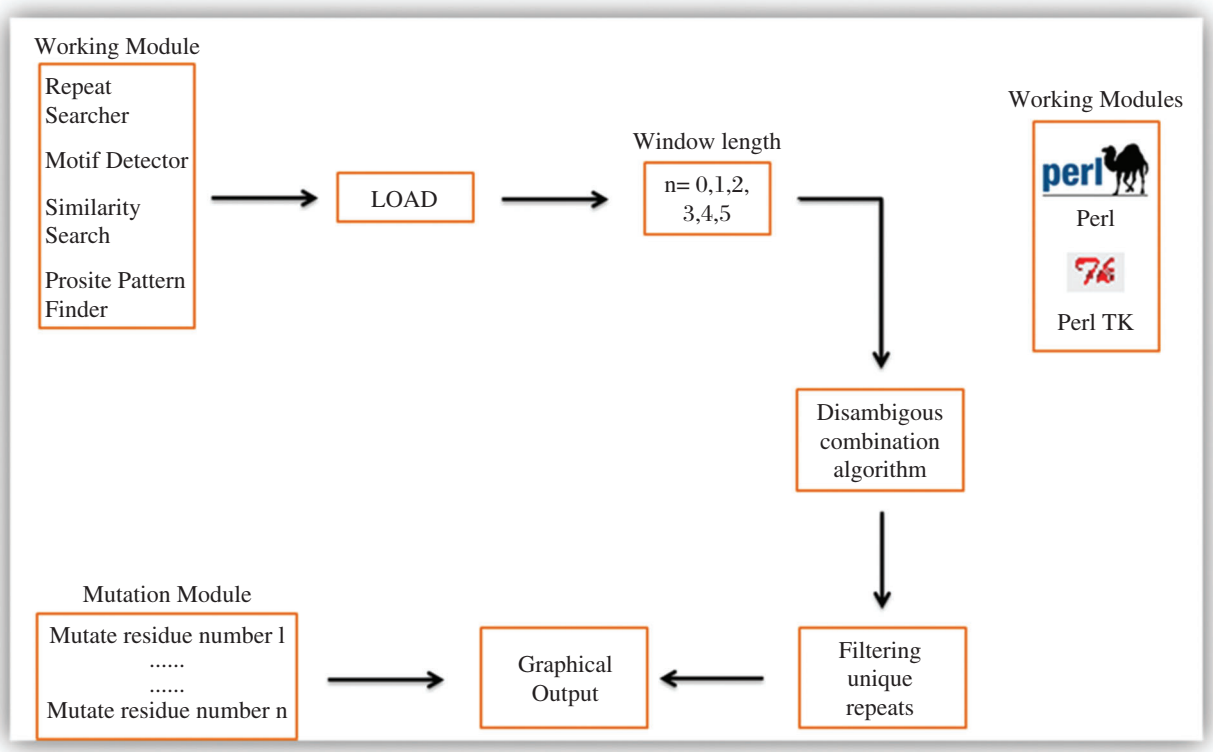
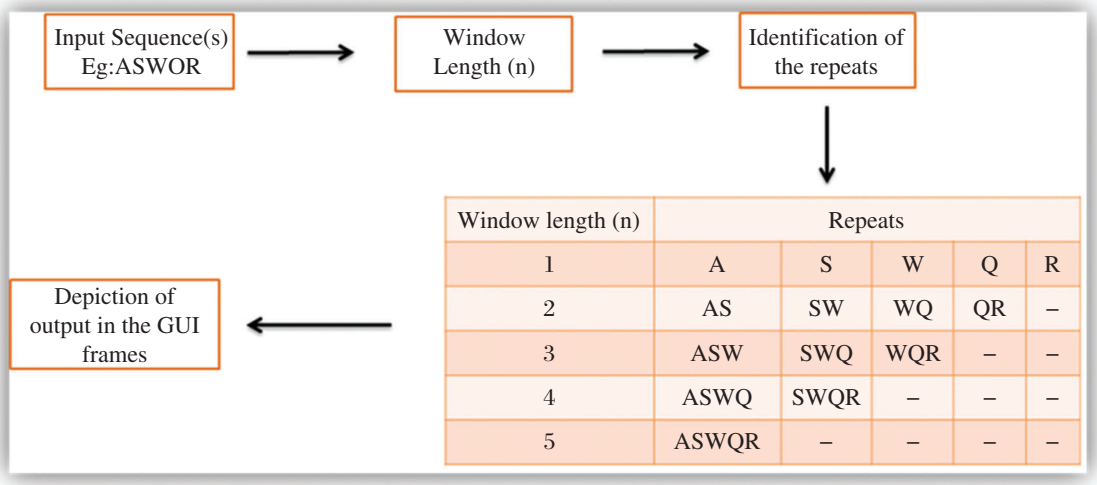


Fig. 2 **Workflow of RSMD**



Window length (n)	Repeats				
1	A	S	W	Q	R
2	AS	SW	WQ	QR	–
3	ASW	SWQ	WQR	–	–
4	ASWQ	SWQR	–	–	–
5	ASWQR	–	–	–	–

Fig. 3 **Workflow of RSMD algorithm**

If the input file contains only one sequence, the algorithm will treat this sequence as the reference sequence. After uploading the input file, the user is made to select a window length. Based on the window size/length, the algorithm generates the repeats from the input sequence (s) which is stored in another array. Unique repeats from this array is selected and displayed in the first GUI frame along with its frequency in the input sequence (s). The second GUI frame displays the different locations of each repeat (***Fig. 4***). The algorithm also identifies the motifs that are present in the input sequence. RSMD contains an inbuilt database of various motifs obtained from different literature sources and other databases like PROSITE. The sequence is matched to all the present motifs in the database and the various motifs present in the sequence are graphically depicted (***Fig. 3***).

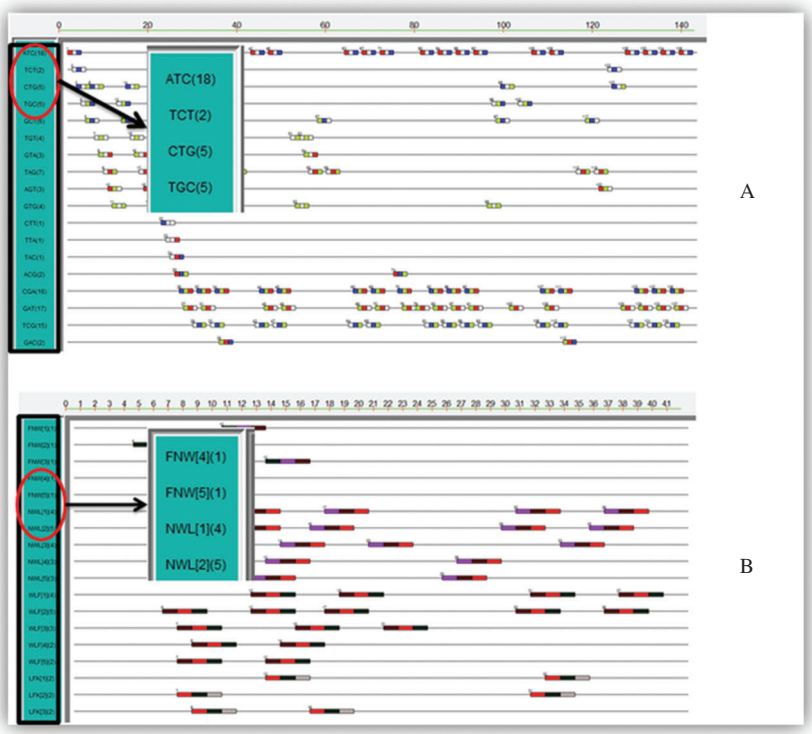


Fig. 4 **Comparison of the graphical output between single sequence and multiple sequences input.** A: Single sequence (format: repeat(frequency)). B: Multiple sequences (format: repeat (frequency)[sequence_num]).

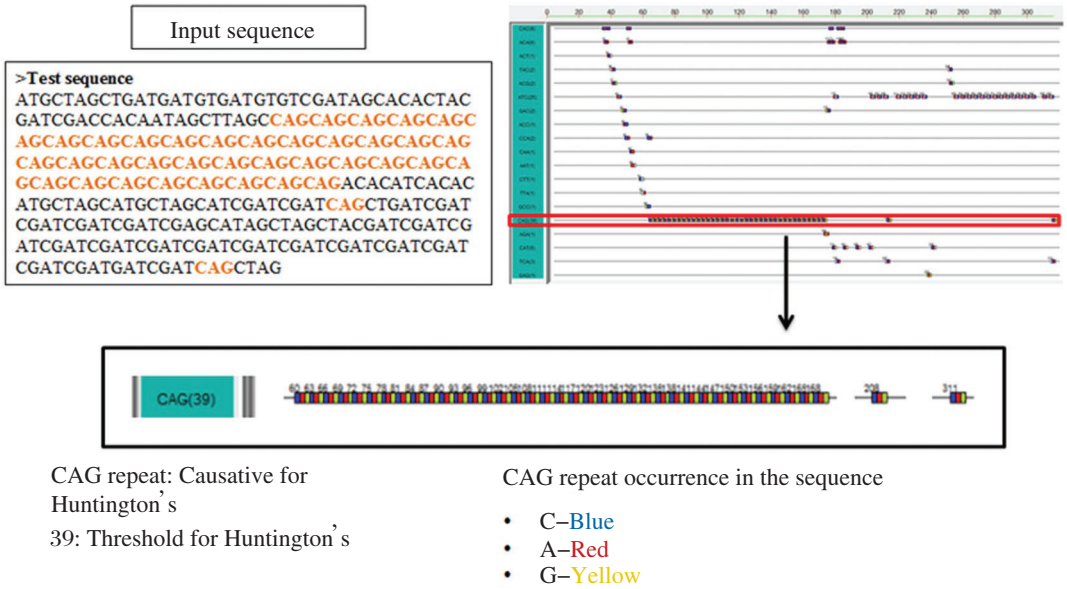


Fig. 5 **Case study depicting the RSMDs capability in identifying disease causing repeats.**

CASE STUDY

A small case study was performed to display RSMDs capability in identifying disease causing repeats. Huntington's disease is a neurological disease that is caused by excessive occurrence of CAG repeats. It is known that Huntington's disease is caused by expansion of the stretch of 36 CAG repeats. A sequence with a frequency of 39 CAG repeats was performed as input to RSMD. The window length which was given as 3 amino acids (CAG) has 3 amino acids in the repeat. The unambiguous combination algorithm generated various repeats and depicted 39 occurring CAG repeats. The output was manually verified with the sequence and was found to be accurate (***Fig. 5***).

CONCLUSION

The significance of the tandem repeats and the volume of research in repeats identification necessitate efficient tools to perform the operation. The graphical uniqueness, computational speed and accuracy of RSMD make it a promising tool, which is highly effective and user friendly. These added features have made the tool a unique platform in bioinformatics research and application that helps researchers reduce the complexity of their tasks in identification of repeats and motifs. The computational performance of RSMD is compared with the existing software and tools to assess its performance (***Table 2***). It is observed that RMSD is significantly faster (***Fig. 6***) and more accurate compared to its competitors. The inbuilt database for storing motifs and the unambiguous combination generator for generating repeats are the chief pillars of RSMD that guarantee the usefulness of the tool. The modularity of Perl has been incorporated to boost the speed of calculation in RSMD and the presentation of the output. The entire development of the tool has been carried out by the fact that biologists with limited computational knowledge consist of the main users of the tool. Therefore, the interface has been designed in a easy-to-use and understand manner. Identification of tandem repeats in particular microsatellites, for large genomic or proteomic sequences is taxing. The relative ease with which RSMD handles this task is anthor outstanding feature of this tool. The further add-ons of RSMD in development include the identification of

Table 2 **Comparison of RSMD with various other existing tools.**

Name	Algorithm used	Flanking region	Repeat finder	Motif identification	Multi sequence testing	User interface
Repeat finder	Repbase	NO	YES	YES	NO	Console/WEB
Find pattern	Perfect imperfect	NO	YES	YES	NO	GUI
TAMO	Combination of above tools	YES	YES	NO	NO	GUI
STAR	Minimum description length	NO	NO	NO	NO	Console
RSMD	Disambigous combination algorithm	YES	YES	YES	YES	GUI

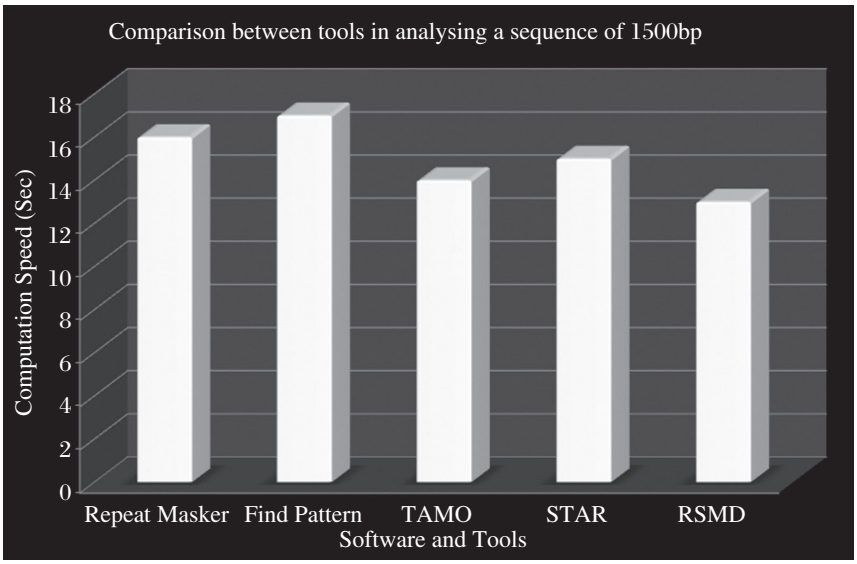


Fig. 6 **Comparison of various tools with RSMD based upon computational speed in seconds.**

all the high repetitive sequences and moderate repetitive sequences. The addition of newly discovered motifs will be added manually. To remove this burden in future, we are trying to incorporate a text mining software by which the data updating is automatic. This comprehensive tool will prove to be valuable in various departments of Life Sciences. RSMD has also been tested in various platforms of bioinformatics and has proven to be very useful in expanding the iota of sequence motif and repeats in protein and DNA sequences. The corresponding author may be contacted for any queries and suggestions about RSMD.

SYSTEM REQUIREMENTS

RSMD was designed to run on Windows operating systems with an installed Perl interpreter. PerlTK is used for providing the GUI. The frame size of the tool may vary depending on the resolution of the system screen. Higher configuration machines are more preferable in large sequence analysis. A minimum of a dual core processor and a 2GB RAM is required to meet the computational speed of RSMD. The tool can be downloaded from the RSMD help center: http://www.bioindians.org/RSMD. A detailed tutorial on RSMDs will be available in the RSMD.rar file.

Acknowledgement

We thank SASTRA University for providing the wonderful infrastructure to carry out our work successfully. We thank our associate dean Dr. M. Vijayalakshmi for her never ending support. We thank Dr. N.T. Saraswathi, Dr. K. Saraboji and Dr. S. Thamotharan for their helpful discussions. We thank Dr. M.V. Satish kumar for commenting on the manuscript.

References

[1] McMurray CT. DNA secondary structure: a common and causative factor for expansion in human disease. *Proc Natl Acad Sci USA* 1999;96:1823–5.

[2] Keniry MA. Quadruplex structures in nucleic acids. *Biopolymers* 2000;56:123–46.

[3] Shafer RH. and Smirnov I. Biological aspects of DNA/RNA quadruplexes. *Biopolymers* 2000;56:209–27.

[4] Reddy PS and Housman DE. The complex pathology of trinucleotide repeats. *Curr Opin Cell Biol* 1997;9:364–72.

[5] Timchenko LT and Caskey CT. Triplet repeat disorders: discussion of molecular mechanisms. *Cell Mol Life Sci* 1999;55;1432–47.

[6] Susan Andrew E, Paul Goldberg Y, Berry Kremer, Hakan Telenius, Jane Theilmann, Shelin Adam, et al. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington′s disease. *Nature Genetics* 1993;4;398–403.

[7] Mudunuri SB, Nagarajaram HA. IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* 2007;23;1181–7.

[8] van Belkum A, Scherer S, van Alphen L and Verbrugh H. Short sequence DNA repeats in prokaryotic genomes. *Mol Biol Rev* 1998;62:275–93.

[9] Scott KD, Eggler P, Seaton G, Rossetto M, Ablett E M, Lee L S, et al. Analysis of SSRs derived from grape ESTs. *Theor Appl Genet* 2000;100:723–6.

[10] Karaca M, Saha S, Jenkins JN, Zipf A, Kohel R and Stelly DM. Simple sequence repeat (SSR) markers linked to the Ligon Lintless (Li1) mutant in cotton. *J Hered* 2002;93:221–4.

[11] Christian Sigrist JA, Lorenzo C, Nicolas H, Alexandre G, Laurent F, Marco P, et al. PROSITE: A documented database using pattern and profiles as motif descriptors. *Brief Bioinform* 2002;3;265–74.

[12] Natalia V, Brain JH and Steven Salzberg. A clustering method for repeat analysis in DNA sequences. *Genome Biology* 2001;2;1–11.

[13] Benjamin GD, Nekludova L, McCallum S and Fraenkel E. TAMO: a flexible, object oriented framework for analysing transcriptional regulating using DNA-sequence motifs. *Bioinformatics* 2005;21;3164–5.

[14] Delgrange O and Rivals E. STAR: an algorithm to Search for Tandem Approximate Repeats. *Bioinformatics* 2004;20;2812–20.

[15] Sarani R, Udayaprakash NA, Subashini R, Mridula P, Yamane T and Sekar K. Large cryptic internal sequences repeats in protein sequences from *Homo sapiens. J Biosci* 2009;34;103–12.

[16] John M.H, Michelle S. Simple sequence repeats in proteins and their significance for network evolution. *Gene* 2005;345;113–8.