**Bioinformation** by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

# Finding Alu in primate genomes with AF-1

**Ravi Shankar[1*], Bhavesh Kataria[1], Mitali Mukerji[2]**

[1]Department of Bioinformatics & Structural Biology, School of Life Sciences & Biotechnology, Indian Institute of Advanced Research, Gandhinagar, Gujarat (India) 382007; [2]Functional Genomics Unit, 604-A, IGIB, CSIR, Mall Road, Delhi (India) 110007; Email: ravish@iiar.res.in; * Corresponding author

received December 26, 2008; accepted January 27, 2009; published February 26, 2009

**Abstract:**
Repetitive sequences occupy more than 40% of the human genome which is much larger compared to the 2% occupied by the coding DNA. Amongst these Alu elements are the second largest class of repeats, occupying nearly 10% of the whole genome. Alus have been implicated in many genomic processes, sometimes giving rise to aberrations while many times playing as silent player in genomic and regulatory evolution. Here we present a web server, AF1, exclusively developed for finding Alu like elements. Besides alignment based methodology, this server utilizes probabilistic scanning to find more diverged elements and employs a more precise way of element classification based on unequal weighting of sequence through sequence encoding.

**Availability:** AF1 is freely available at: http://software.iiar.res.in/af1/. The standalone is also available for download.

**Keywords:** Alu; algorithm; non-coding; primate; repetitive element

**Background:**
Alus are short interspersed nucleotide elements, which comprise about 10% of human genome (International Human genome consortium, 2001). An intact Alu has two monomeric units linked through an A-rich region, with approximately 67% identity. The average length of these elements is estimated to be 282 base pairs excluding variable length 3' poly-A tail. These repeats harbor regulatory sites and contribute to the regulatory repertoire of the genome **[1, 2]**. Alu repeats have been implicated in alternative splicing and coding for proteins **[3]**. Many monogenic diseases like acute myelocytic leukemia, Tay Sach's and hemophilia are associated with Alu transpositions **[4]**. Alu repeats have been a very useful marker in phylogenetics and evolution based studies **[5, 6]**.

So far two generalized repeat finding programs have been extensively used for Alu finding: RepeatMasker and Censor **[7]**. These programs have BLAST in their core and find broad spectrum of repeats like complex as well as simple repeats. Both the programs are dependent upon a common database, Repbase [**8**]. Here we have tried to find Alu kind of elements by implementing an alignment based method in combination with a probabilistic modeling which utilizes matrices specially designed on Alu sequences, to analyze Alu elements in the ever increasing amount of primate sequences and interest in non-coding genomic sequences, majority of which earlier tagged as junk. Another important feature is its classification, which incorporates unequal weighting of positions to minimize the impact of non-diagnostic position in determining the class. Also AF1 would be first of its kind as an exclusively dedicated server for Alu elements.

**Input and output:**
The basic working principle of AF1 is shown in **Figure 1**. The AF1 server has mainly following components: (1) Restricted alignment based module (2) Probabilistic modeler (3) Classifier. The user gives an input sequence in FASTA format either in paste sequence mode or load sequence file mode. The server has been designed to take a large single sequence. The input query sequence is searched for exact match seed using library of overlapping words generated from an Alu prototype sequence unlike any other database search tools that instead break query into words to scan databases. For every hit only flanking 300 bp regions on both ends are taken for further analysis through alignment. These subsequences are subjected to first scan for longest possible region of continuous match to nucleate the alignment. Unlike other famous methods of detecting multiple nucleus, here we need to locate just one and around which alignment is extended. The matrices used are specific for Alu, derived from 5000 Alu sequences. If Alu is not detected by this alignment, the entire alignment is scanned for a small subregion having reasonable identity. If its present, the aligning sequence is subjected to probabilistic scanning where PWM derived from alignment of 5000 Alu sequences is used with overlapping window of 32 on matrix as well as on sequence, assuming each position as start position in the matrix as well as sequence. The score is compared to random one using a randomized matrix with same dimensions and composition and evaluated for threshold value for identification as an Alu. The found Alu repeats are presented in both directions, whose links are made available. Clicking on those links provides tabulated results giving start and end position with found Alu in that region. Probabilistic approaches work well when sequences are not very close and in case of Alus when they are old and highly diverged.

The last stage is classification where the query is converted into encoded sequence via alignment with Alu Sx



ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 287-288 (2009)

prototype. The same is done for all known subfamilies of Alus. Finally the encoded query is aligned to encoded subfamily library where only diagnostic position is allowed to guide the alignment and achieve the correct judgment for classification. Classification option runs automatically once the first step of Alu identification is complete. The output of classification step is start of the region, end of the region, classified subfamily and sequence.

The entire server has been implemented in Tomcat with JSP, while the core programs have been written in C++, Python and PERL. Details, comparison and algorithm of program are available on the server page. The program achieved sensitivity and specificity above 0.9 when validated over experimental data from various sources. This data too has been made available on the server.

**Caveats and future development:**

The web server version has some limitations with size of query as it takes some amount of time if query is very large. It is our server limitation which we are trying to fix by converting our code for parallel computing and run the server through 64 nodes cluster. A possible issue could be the time taken in classification step. This part too could be made faster in the future. Presently we have made the standalone version of the software available on download section which users can easily install on their systems. More complex models can be incorporated for probabilistic scanning in future to get better result for highly weathered elements as well as the entire methodology can be extended to other transposons or retroelements. Continuous work over the server will keep on going in order to keep it up-to-date and refined.

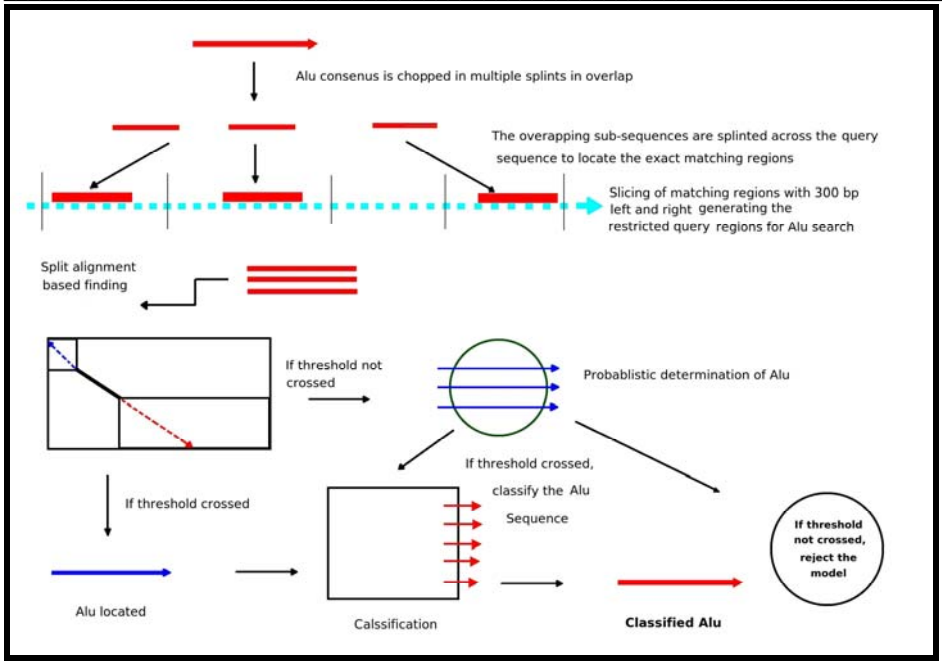


**Figure 1:** AF-1 working principle

**Acknowledgments**

Initial support of the project was obtained from IGIB, CSIR. We thank Amit Chaurasia at IGIB for testing support. We are thankful to the Puri Foundation for its support for funding and human resources. We are also thankful to the sub-DIC Bioinformatics Center at Indian Institute of Advanced Research, Gandhi Nagar for its support.

**References**

**[1]** R. Shankar, *et al.* , *BMC Evol Biol* 2004, **4:** 37 (2004) [PMID: 15461819]

**[2]** N.V. Tomilin, *et al.*, *Nucleic Acids Res*, **20:** 2941 (1992) [PMID: 1620588]

**[3]** W. Makalowski, *et al.*, *Trends Genet*, **10:** 188 (1994) [PMID: 8073532]

**[4]** P. L. Deininger & M. A. Batzer, ,*Mol Genet Metab*, **67:** 183 (1999) [PMID: 10381326]

**[5]** R. Shankar, *et al.*, *Mol Genet Genomics*, **277:** 441(2007) [PMID: 17375324]

**[6]** H. Hamdi, *et al.*, *J Mol Biol*, **289:** 861 (1999) [PMID: 10369767]

**[7]** J. Jurka, *et al.* ,*Comput Chem* 1996, **20:** 119 (1996) [PMID: 8867843]

**[8]** J. Jurka, *et al.*, *Cytogenet Genome Res*, **110:** 462 (2005) [PMID: 16093699]

**Edited by P. Kangueane**

**Citation: Shankar *et al.***, Bioinformation 3(7): 287-288 (2009)





ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 287-288 (2009)