Abstract

The whole genome has just ~5% of coding DNA, more than 40% of the genome is composed of repetitive sequences. Alu repeats are the second largest repeats, composing about 10% of the whole genome, considering the result of presently available repeat finding softwares. Earlier considered as junk DNA having no significant biological function, the recent studies have demanded more attention on these repeats where possibility of biological roles being played by these elements have been emphasized. Besides this Alus are very significant component for evolutionary and phylogenetic studies.

Alu repeats hold their own importance in the ongoing genomic studies. Finding Alu repeats in the genome is an essential part of these studies. One needs to depend upon the annotation softwares presently available like CENSOR and RepeatMasker. These softwares have been widely used to annotate repeats in the genomic sequences of primates. There are chances to get more repeats in the genomic sequences not earlier reported by the methodology used by any of these two programs. Here we present an exclusively designed Alu finding program AFl(AluFinder-1), which is able to find Alu repeats successfully from the DNA sequences in pre annotated dataset as well as 33 ENCODE regions. The program uses a set of overlapping patterns generated from Alu sequences to restrict the operation area followed by three step alignment method, position weight matrix generated from 4000 aligned Alu repeats and subsequent overlapping scan within the matrix as well as query. Compressed pattern alignment method to find the closest class of the Alu subfamily is used to classify the found sequences. Biological informations exclusive to Alu repeats have been used in the program. We expect AFl to be useful in finding Alu repeats successfully in the primate genomes. The web server version of the program is available at http://203.90.127.77/af1.html.
Non random genomic divergence in repetitive sequences of human and chimpanzee in genes of different functional categories

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Molecular Genetics and Genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>MGG-06-00128</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Original Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>18-Jul-2006</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Shankar, Ravi; IGIB, GNRKC genome informatics lab  
|                     | Chaurasia, Amit; IGIB, GNRKC genome informatics lab  
|                     | Ghosh, Biswaroop; IGIB, GNRKC genome informatics lab  
|                     | Chekmenev, Dmitry; BIOBASE GmbH, R & D  
|                     | Cheremushkin, Evgeny; A.P. Ershov's Institute of Informatics Systems, R & D  
|                     | Kel, Alexander; BIOBASE GmbH, R & D  
|                     | Mukerji, Mitali; IGIB, GNRKC genome informatics lab  |
| Keywords:      | Genome, Repeat, Neuronal, Evolution, Human |
Non random genomic divergence in repetitive sequences of human and chimpanzee in genes of different functional categories

Ravi Shankar¹, Amit Chaurasia¹, Biswaroop Ghosh¹, Dmitry Chekmenev², Evgeny Cheremushkin³, Alexander Kel³, Mitali Mukerji¹*

¹Functional Genomics Unit, Institute of Genomics and Integrative Biology, CSIR, Mall Road, Delhi, India.
²BIOBASE GmbH, Halchtersche Strasse 33 D-38304 Wolfenbuettel, Germany.
³A.P. Ershov's Institute of Informatics Systems, 6, Lavrentiev ave., 630090 Novosibirsk, Russia.

Emails:
Ravi Shankar (ravis@igib.res.in);
Amit Chaurasia (achaurasia@igib.res.in);
Biswaroop Ghosh (brghosh@igib.res.in);
Dmitry Chekmenev (dmitri.chekmenev@biobase-international.com);
Evgeny Cheremushkin (cher@biobase-international.com);
Alexander Kel (alexander.kel@biobase-international.com);

*Correspondence may be addressed to:
Dr. Mitali Mukerji
Institute of Genomics and Integrative Biology, CSIR,
Mall Road,
Delhi 110007, India.
Tel: 91-11-27667602
Fax: 91-11-27667471
E-mail: mitali@igib.res.in

Keywords: Evolution; Repeats; divergence; transcription; Genome; Genes; genetics; Human; chimpanzee; Neuronal
Abstract
Sequencing of the human and chimpanzee genomes has revealed ~99% similarity in the coding sequence between both the species, which in no way parallels the observable phenotypic differences. In this context, we have explored the contribution of non-coding repetitive DNA in functional divergence between human and chimpanzee. This study has been carried out in a dataset of neuronal and housekeeping genes whose coding regions have been earlier extensively compared between the two species.

Comparative analysis of dinucleotide content density divergence, repeat spectrum, divergence and partitioning of divergence between repeats and transcription factor binding sites differences, indicate different extent of functional constraints associated with the non-coding repeat regions. The neurodevelopmental genes seem to diverge more in the intronic region whereas the neurophysiology genes show higher divergence in the upstream 2kb region. Most of the divergence observed in the housekeeping genes is contributed by repeats. We also observe gain of function specific transcription factor profiles in the human lineage. Interestingly, a major fraction of these function specific sites is differently partitioned in the repetitive sequences present in these regions depending upon the relative distribution of the repeats across the functional categories.

Differential distribution of repeats across the various functional categories could impact genome wide regulation and structure to a large extent. The insights obtained from this study further add a new facet to the contribution of non-coding factors in evolution.

Introduction
Genome sequence analysis of chimpanzee and human has revealed an average divergence of about 1.2 % between the two species(Chen and Li, 2001; Chen et al., 2001). This observation is difficult to reconcile with, when one compares the morphological, cognitive or behavioral skills between both the species(Cheng et al., 2005; Dorus et al., 2004; Watanabe et al., 2004; Caceres et al., 2003; Gilad et al., 2003; Gu and Gu, 2003; Krubitzer and Kahn, 2003; Britten, 2002; Enard et al., 2002; Chen and Li, 2001; Chen et al., 2001). The vast amount of variability and diversity both within and between human and chimpanzees(Britten, 2002; Chen and Li, 2001; Chen et al., 2001) make it even more challenging to distinguish variations which mark speciation events from those which determine intra-species variability. Comparative analysis of the recently available sequence of the entire chimpanzee genome with human has revealed many regional variations in the genomic landscape(Cheng et al., 2005). Though the average sequence