Poster H-52

Web Based Computational Tools for Predicting G-quadruplexes in Mammalian Genes

Authors:
Oleg Kikin (*Ramapo College of NJ*)
Rumen Kostadinov (*University of Pennsylvania*)
Nishtha Malhotra (*Ramapo College of NJ*)
Manuel Viotti (*Ramapo College of NJ*)
Lawrence DAntonio (*Ramapo College of NJ*)
Paramjeet S. Bagga (*Ramapo College of NJ*)

Short Abstract: Highly stable G-quadruplex structures formed by guanine rich nucleic acids have recently come into limelight because of their potential role in biologically important processes, including regulated RNA processing. We have developed a suite of web based computational tools for predicting G-quadruplexes and studying their distribution patterns in mammalian genes.

Long Abstract:

The quadruplex structures formed by guanine rich nucleic acid sequences have received significant attention recently because of increasing evidence for their role in important biological processes and as therapeutic targets. The G-quadruplex structure is formed by repeated folding of either the single polynucleotide molecule or by association of two or four molecules. The structure consists of stacked G-tetrads, which are square co-planar arrays of four guanine bases each. G-quadruplex is stabilized with cyclic Hoogsteen hydrogen bonding between the four guanines within each tetrad.

G-quadruplex sequence motifs have been reported in telomeric, promoter and other regions of mammalian genomes. G-quadruplex DNA has been suggested to regulate DNA replication and may control cellular proliferation. Although initially most of the studies focused on G-quadruplexes in the DNA, lately there have been many efforts to study G-quadruplex forming RNA. In fact, G- rich sequences capable of forming G-quadruplexes in the RNA have been implicated in a variety of important biological activities, such as mRNA turnover , Fragile X Mental Retardation Protein (FMRP) binding, translation initiation as well as repression.

We have previously shown that a conserved auxiliary G-rich sequence (GRS) found near the polyadenylation regions can mediate efficient 3' end processing of mammalian pre-mRNAs by interacting with DSEF1/hnRNP H/H' protein. Regulated polyadenylation is an important component of differential gene expression. An interplay among GRS-binding proteins helps in regulating alternative polyadenylation of mammalian pre-mRNAs. Members of the hnRNP H protein subfamily, that bind G-rich motifs, are also known to be involved in alternative, tissue-specific, regulated splicing events. GRS motifs that are present near splice sites act as splicing regulators by interacting with hnRNP H. The regulatory G-rich motifs may be capable of forming quadruplex structures. Whether quadruplex structure directly plays a role in regulating RNA processing events requires investigation.

The majority of the mammalian poly(A) region GRS sequences that we had surveyed in our previous studies are capable of forming unimolecular G-quadruplexes. Our preliminary

analysis of alternatively processed human transcripts has also revealed the presence of quadruplex forming sequences near alternative splice sites. However, a more detailed investigation into the distribution of G-quadruplex sequences near RNA processing sites requires a systematic large-scale analysis of mammalian genes.

We have used a computational approach to map putative Quadruplex forming G Rich Sequences (QGRS) within the transcribed regions of a large number of alternatively processed human and mouse genes. Our computational suite consists of a “QGRS Mapper” program that can analyze fully annotated genomic nucleotide sequences from NCBI-based databases, and the “GRSDB” database for curation and further analysis of QGRS Mapper data on alternatively spliced and alternatively polyadenylated human and mouse genes.

QGRS Mapper (http://bioinformatics.ramapo.edu/QGRS/) is capable of analyzing user-provided RNA/DNA sequences in the raw or FASTA format for the analysis of biologically important genomic segments, e.g. promoter and telomeric regions. It is also useful for predicting G-qudruplex structures in oligonucleotides. However, the main function of the program is to search and retrieve desired gene/nucleotide sequence entries from the NCBI databases for mapping putative G-quadruplexes (QGRS) in the context of RNA processing sites. One can search and analyze gene sequences by Gene ID, Gene Name or Symbol, Accession number or GI number for an NCBI nucleotide sequence entry. Once appropriate genes have been identified, this program links to GenBank or RefSeq, downloads the corresponding genomic nucleotide sequence entry of the gene, and parses the entry for product, intron, exon, polyA and related information. The program then processes the nucleotide sequence to find all QGRS and map their location within the gene and their distance from relevant RNA processing sites. Options are provided that allow the user to specify the length of the QGRS and structural elements such as the number of guanine tetrads and the size and composition of the intervening loops. We have also devised a scoring system that evaluates a QGRS for its likelihood to form a stable G-quadruplex. The method is based on well documented principles that have established a relationship between the stability of a G-quadruplex with the number of guanine tetrads and its loop lengths.

In addition to providing data on composition and locations of QGRS relative to the processing sites in the pre-mRNA sequence, QGRS Mapper features an interactive graphic representation of the data. The interactive graphics module can be used to visualize QGRS distribution patterns among all the alternative RNA products of a gene simultaneously. The graphics view allows the user to zoom in on QGRS in any portion of the nucleotide sequence.

We have been using a specialized version of the QGRS Mapper to analyze alternatively processed (alternatively spliced or alternatively polyadenylated) mammalian genes. The results of these analyses are uploaded into GRSDB, a relational database built using MySQL. GRSDB (http://bioinformatics.ramapo.edu/grsdb/) is specially structured to facilitate queries about alternatively processed genes. While GRSDB currently contains data only on human and mouse genes, our computational tools and the database are designed to include other organisms as well. GRSDB is continuously being updated with new data entries.

Researchers interested in evaluating the ability of nucleotide sequences to form unimolecular G-quadruplexes will find QGRS Mapper and GRSDB to be very useful tools. Due to the flexible and comprehensive nature of the design, these programs are expected to

serve a variety of scientists. The applications will be especially attractive to individuals interested in exploring the role of G-quadruplexes in regulated RNA processing.

We have been using these servers to perform a large scale analysis of alternatively processed mammalian transcripts. At present, our database contains information obtained from 1310 human and mouse genes, of which 1188 are alternatively processed. A total of 30 584 introns and 33 816 exons were analyzed, containing a total of 3231 RNA products. Almost all of the transcripts exhibited the presence of G-quadruplex elements near splice sites and poly(A) regions. These elements were selectively associated with processing sites of alternate gene products. Our findings lend supporting evidence that G-quadruplex elements could play a regulatory role in differential RNA-processing.