Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 47(2) 321-326 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

QuadLocator - An Algorithm to Locate Possible Quadruplex Forming Sequences

F. N. MEMON[++], Z. U. A. KHUHRO, A. P. HARRISON*

Institute of Mathematics and Computer Science, University of Sindh, Jamshoro

Received 2nd January 2015 and Revised 20th May 2015

Abstract— *An unusual four stranded structure, known as G-quadruplex structure, can be formed in a DNA/RNA segment having frequent guanines. These structures can be involved in different biological processes and to understand this, it is necessary to identify the locations of those segments that can form these unusual structures. Different tools are available to locate such sequences in a given nucleic acid sequence. However, a new algorithm is proposed in this paper with an additional feature that can perform transcript-exon mapping for these structures. It is important because there is a possibility that a possible quadruplex forming sequence may found because of joining of two exons in a mature transcript but not in case of DNA or RNA.*

Keywords- Quadruplex structure, DNA, RNA, exon-exon junction, transcript-exon mapping

1. INTRODUCTION

A guanine rich DNA/RNA segments can form unusual four stranded structures, known as G-Quadruplex structures. The details of their structure and formation are available elsewhere (Huppert 2005, Karimata 2005, Bates 2007, Armitage 2007). However, Qin in 2008 mentioned that the identification of such sequences helps identifying the biological role of G-quadruplex structures. Zahler (1991), Darnell (2001) and Siddiqui-Jain (2002) suggested that G-quadruplex structures may be involved in various biological processes including translational control, regulating telomerase activity, and transcriptional repression. Recent advances have demonstrated that quadruplex structures can play a role in gene expression and provide opportunities for a new class of anticancer therapeutics and drug targets (Qin 2008, Yadav 2008).

All these facts show the importance of deep studies of genomic and transcriptomic data in order to find the biological impact of G-quadruplex structures. To carry out such a study, it is necessary to locate the possible G-quadruplex forming sequences in an organism's genome. The availability of tools to identify quadruplex forming sequences can help the researchers to find the genomic/transcriptomic regions for possible G-quadruplex forming sequences which can then help to look at their role in biological processes.

Due to these unusual structures, the problems are also found in microarray technology that is used by scientists for gene expression measurements and other purposes. Our previous work has presented some of these problems (Upton 2009) whereas we have further focused on G-Quadruplex formation on microarray specifically on Affymetrix GeneChips (Upton 2008, Langdon 2009, Memon 2010a, Memon 2010b, Shanahan 2012). Wu *et al.* have also presented the abnormal effects of short sequences containing continuous guanines (Wu 2007) and it is assumed that these effects are due to the formation of G-quadruplex structures on the surface of Affymetrix microarrays.

This paper introduces an algorithm, the QuadLocator that has been designed to find the guanine rich sequences that are capable of forming G-quadruplex structures in nucleic acid sequences. Many next generation sequencing platforms are measuring ssDNA/RNA and this new technology is replacing the most dominant Sanger sequencing technology. In future, the QuadLocator will also be beneficial to analyze the next generation sequencing data for the possible quadruplex forming sequences (PQFSs).

There are a few algorithms that have been designed to identify the quadruplex forming sequences. Quad-parser (Huppert 2005), Quadfinder (Scaria 2006), QuadBase (Yadav 2008) are some examples of these algorithms. Some of these methods do not accept a nucleic acid sequence as an input file. For a long sequence such as an Ensembl chromosome sequence, it is difficult to type the sequence in the given text box. In another case, if the downloaded Ensembl chromosome sequence is copied in the text box (given for input sequence), some method results in an error because Ensembl chromosome sequences are available in FASTA format in which the sequence is broken down and each line of the file contains 60 bases and a new

[++]Corresponding author: F. N. Memon, Email: farhatnm@usindh.edu.pk
*Department of Mathematical Sciences, University of Essex, UK

line characters. So the existing algorithms that do not accept new line character will not work in this situation. In this case, a FASTA file needs to be modified where all new line character should be removed either manually that takes time or automatically that needs a script. Secondly, if an input sequence is broken into more than one line, one of the existing algorithms does not support it to find PQFSs successfully.

Another reason to develop such an algorithm is that besides finding the locations of PQFS, the QuadLocator has an additional feature to facilitate its users by performing the PQFS mapping among the transcripts and their exons. The main algorithm, QuadLocator, finds PQFSs in a single stranded sequence that could be a DNA sequence, a RNA sequence (primary or mature transcripts), an exonic sequence, or any other sequence of similar type. However, this extra feature of the QuadLocator is designed to map the PQFSs among the transcripts and their exons.

A DNA sequence or RNA sequence (primary transcript) contains introns in between the exonic regions. These introns are removed during the synthesis of a mature transcript. Thus, a DNA sequence or its equivalent RNA sequence for a particular gene is different than the sequence of that gene as a mature transcript (messenger RNA) as the latter does not contains introns. It is therefore possible that the PQFSs are different in these two cases for that gene. The following illustrations are showing some of the examples of these differences.

Illustration 1:

Figure 1 is showing an example of DNA sequence of an unreal gene. It also demonstrates the sequences of primary and mature transcripts. The differences of the three sequences can be seen in the figure. In short, a primary transcript is similar to the DNA sequence of a gene with a difference that all the thymine bases are replaced by uracil bases. Then this primary transcript is used to synthesize the mature transcript in which intronic regions are removed.

The PQFSs in DNA and primary transcript are similar due to the fact that the two sequences are same accept with the uracil and thymine replacement **(Table 1).** However the sequence of mature transcript may have different PQFSs depends on the deleted intronic regions. Table 1 shows three PQFSs for DNA and primary transcript while only one PQFS for the mature transcript. The reduction in PQFSs in the mature transcript is the consequence of removing introns.

Table 1: List of possible quadruplex forming sequences for the unreal gene given in Fig. 1.

Pattern	Start	End	Length	Sequence
PQFSs for DNA				
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	2	26	25	GGGTCGGGCCATGGGCTTTAGGGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	7	33	27	GGGCCATGGGCTTTAGGGGGAAGGGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	14	42	29	GGGCTTTAGGGGGAAGGGGGCGGTGGGGG
PQFSs for Primary Transcript				
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	2	26	25	GGGUCGGGCCAUGGGCUUUAGGGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	7	33	27	GGGCCAUGGGCUUUAGGGGGAAGGGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	14	42	29	GGGCUUUAGGGGGAAGGGGGCGGUGGGGG
PQFSs for Mature Transcript				
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	2	28	27	GGGUCGGGCUUUAGGGGGAGGGUGGGGG

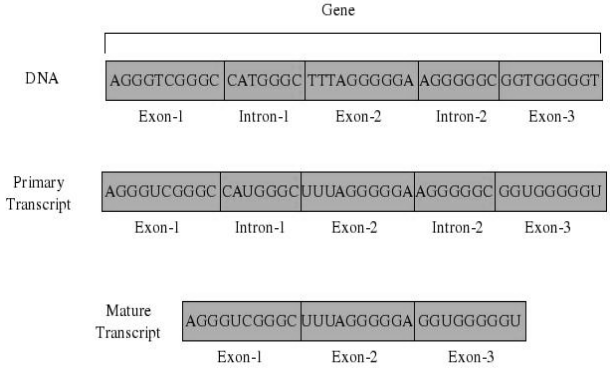

Fig. 1: An example of DNA sequence, primary transcript, and mature transcript of an unreal gene to show the possible quadruplex forming sequences (given in Table 1and discussed in illustration 1).

Illustration 2:

Fig. 2 is showing another example of DNA sequence of a different unreal gene and the PQFSs of the DNA, primary transcript, and mature transcript are given in **Table 2.**

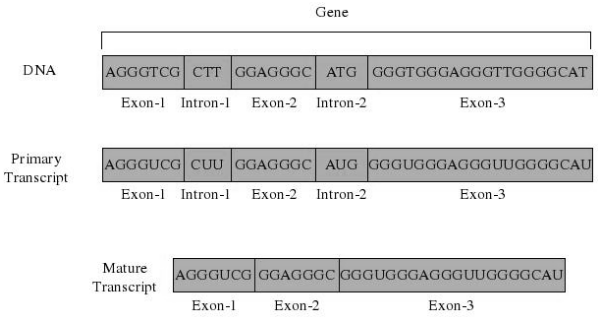

Fig. 2: An example of DNA sequence, primary transcript, and mature transcript of another unreal gene to show the possible quadruplex forming sequences (given in Table 2 and discussed in illustration 2).

Table 2: List of possible quadruplex forming sequences for the unreal gene given in Fig. 2.

Pattern	Start	End	Length	Sequence
PQFSs for DNA				
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	14	31	18	GGGCATGGGGTGGGAGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	20	37	18	GGGGTGGGAGGGTTGGGG
PQFSs for Primary Transcript				
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	14	31	18	GGGCAUGGGGUGGGAGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	20	37	18	GGGGUGGGAGGGUUGGGG
PQFSs for Mature Transcript				
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	2	17	16	GGGUCGGGAGGGCGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	7	21	15	GGGAGGGCGGGUGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	11	25	15	GGGCGGGUGGGAGGG
$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$	15	31	17	GGGUGGGAGGGUUGGGG

Illustration 1 is showing that the removal of introns can reduce the number of PQFSs whereas an opposite case is demonstrated in illustration 2 where the number of PQFSs of mature transcript is increased after removal of introns; four PQFSs for the mature transcript and two PQFSs for DNA and primary transcript.

Thus, the proposed algorithm can not only perform simple search for PQFSs in a nucleic acid or similar type of sequences but it can also perform mapping amongst the transcripts and their exons for PQFS. First we will discuss the main algorithm, the QuadLocator and then the mapping among the transcripts and their exons for PQFS will be presented.

2. METHOD

2.1 QuadLocator - A presented approach

The QuadLocator searches for the sequences of the form $G_{x1-x2}N_{y1-y2}G_{x1-x2}N_{y1-y2}G_{x1-x2}N_{y1-y2}$ where x1 and x2 are the lower and upper limits of number of Gs in a G-run whereas the y1 and y2 are lower and upper limits of the lengths of three loops in between four G-runs. A sequence in this form is considered as quadruplex forming sequences and the proposed algorithm searches for possible quadruplex forming sequences (PQFS) of this form. The QuadLocator is implemented in C and can analyze DNA, RNA and similar type of nucleic acid sequences to locate the PQFSs.

QuadLocator finds sequences with no discontinuities in the stacks of guanines. This is because the effects of guanine replacements/deletions in G-tetrads in order to analyze the discontinuities in the guanine bases show significantly lower stability (Huppert 2005). The algorithm considers a maximum limit of a loop length of 7. Loop lengths 1-7 help to form stable quadruplex structures and an increasing loop length deceases the stability (Huppert 2005). QuadLocator avoids overlapping sequences in such a way that any particular G-run is not considered as the first G-run of more than one sequence.

2.2 How QuadLocator works

Required parameters: The QuadLocator requires the following parameters to work.

(i) (i) Input sequence: A file contains a single stranded nucleic acid molecular sequence (either DNA or RNA). It is known that DNA sequence is composed of a number of four bases: guanines, cytosines, adenine, and thymine. However, in RNA sequence, a uracil takes place of a thymine. Due to this fact, the QuadLocator stops working if it finds an input sequence containing two bases together, thymine and uracil. It also checks if there is any other invalid character in the input sequence.

(ii) Minimum & maximum limits of G-run lengths: Numeric values that must be greater than 1.

(iii) Minimum & maximum limits of loop lengths: Numeric values that must be in the range of 1 and 7.

(iv) Type of quadruplex: One of the characters G, C, T, A. Besides G-quadruplex, the QuadLocator is able to search for C-, T-, or A-quadruplex, if anyone is

interested.

2.3 Algorithm

The QuadLocator performs the following steps to find the PQFSs.

(i) Checks for the limits of G-runs and loops lengths to be in the range. It also checks if the lower limits of the G-run or loop length is greater than upper limits.

(ii) Reads the sequence from an input file and counts the number of each base in the given sequence. These counts are the part of output data.

(iii) Picks the first four G-runs to start the loop.

(iv) Checks how many currently held G-runs are required for a quadruplex and increase the counter if a quadruplex forming sequence can be formed by any number of these G-runs under the limitation of G-run and loop lengths.

(v) Replaces the G-runs in such a way that the second G-run is considered as the first, the third G-run is now considered as second and the fourth as third G-run. Thus, another G-run is required that will be considered as the fourth G-run.

(vi) Reads the next G-run in the input sequence as a fourth G-run.

(vii) Repeats the process to determine another quadruplex forming sequence (from step 4 onwards) using the new selection of four G-runs.

Data structure selected for QuadLocator

A data structure is a specialized way of storing and organizing data that helps to provide an efficient way to access and work with this data. A number of data structures are available that include arrays, linked-lists, trees, files, tables, stacks, queues, and so on. Each data structure has its own characteristics and is suitable for specific applications. An appropriate selection of data structure for an application/algorithm can improve the performance of that algorithm.

The QuadLocator is mainly based on an array that stores the given nucleic acid molecular sequence. The array is found to be most useful because one of its characteristics is that any element X of an array is directly accessible, without accessing other members which are stored before X, if the position/location of X is given. The QuadLocator needs four G-runs at a time to find a quadruplex. The four G-runs are represented by their start and end positions which are provided to the algorithm and it performs its activities using these start and end positions. Using these numbers along with the minimum and maximum limits of G-runs and loops lengths, the algorithm decides which segment is of interest.

2.4 Method for mapping among the transcripts and their exons

The mapping is performed by the following three inter-related steps in such a way that the output of one step will be input for the other.

Step 1:

The first step uses QuadLocator to find the PQFSs for each exon of an organism given in an exon sequence file (a file that contains sequences of all exons of an organism). An output file is generated that contains a list of exon IDs along with the number of nucleotides and number of PQFSs in the exons.

Step 2:

The procedure in step 1 is then repeated for the transcript sequence file (a file that contains sequences of all transcripts of an organism) to get the second output file with a list of transcript IDs along with the number of nucleotides and the number of PQFSs in the transcripts.

Step 3:

The output files of step 1 and step 2 are the inputs of this step. This step performs the mapping among the number of PQFSs in each transcript and its exons. It uses the Transcript-Exon map file (a file that contains information regarding the mapping between transcripts and their exons) and the two files generated in steps one and two. It provides the output in such a way that each line contains the ID of a particular transcript along with the number of PQFSs in that transcript and the collective number of PQFSs in all the exons that make up that transcript. All the input files in any of the three steps can easily be downloaded from Ensembl website using the BioMart.

A transcript may have a number of PQFSs which could be equal to or greater than the number of PQFSs in all its exons. It is noticed that the number of PQFSs in a transcript can be greater than the collective number of PQFSs in all the exons of a transcript. It happens because there is a possibility to find quadruplex forming sequence at exon-exon junction. An exon-exon junction is the place where the sequences of two exons join together. Thus the joining of two sequences can cause the formation of a quadruplex forming sequence. An example of PQFS at exon-exon junction is illustrated in Fig. 2 in which the mature transcript shows no PQFS in exon-1 and exon-2 but one PQFS in exon-3. However, after joining the three exons, four PQFSs were found (see PQFSs for mature transcript in Table 2). The first two PQFSs of mature transcript in Table 2 are formed by joining the three exons, the third PQFS is formed at exon-2 and exon-3 junction and the last one is formed in exon-3.

3. RESULTS AND DISCUSSION

The QuadLocator is applied to the Drosophila melanogaster Genome (Fruit Fly) as an example. The Drosophila Genome is selected because during the

analysis of various GeneChips (Memon 2010b), it was found that one of the Drosophila GeneChip design has a very small number of probes with Continuous guanines, only 64 probes. The lack of high density of such probes could either be a careful selection of probes or the Drosophila genome can have fewer tendencies to form G-quadruplex structures. Hence the DNA sequences of Drosophila genome were downloaded from the Ensembl website (www.ensembl.org/index.html). The entire genomic sequence is broken up into several files where each file contains the sequence of a chromosome. Table 3 shows the results of QuadLocator when applied on Drosophila chromosome sequences. For these particular results, the size of G-runs is chosen between 3 and 5; whereas the length of loops is considered in the range of 1 to 7. The genomic sequence is analyzed for G-quadruplex forming sequences.

Table 3: List of Drosophila chromosomes along with their sizes and number of PQFSs. All the Drosophila chromosome sequence files were downloaded in FASTA format from Ensembl website on October 05, 2011.

Chromosome	No. of bases	No. of G-quadruplexes
Chromosome 2L	23,011,544	51,745
Chromosome 2R	21,146,708	53,683
Chromosome 3L	24,543,557	57,588
Chromosome 3R	27,905,053	69,954
Chromosome 4	1,351,857	1,071
Chromosome X	22,422,827	67,394

The QuadLocator took approximately 22 seconds to find the PQFSs for the entire Drosophila genome on a machine with 2.3 GHz of speed. The QuadLocator is also able to show the exact location of PQFSs.

4. CONCLUSION

If a nucleic acid sequence has frequent occurrences of guanines, it is able to form some four stranded structures that are called G-quadruplex structures. Many researchers are interested to identify the biological role of these structures for which they need to know the locations of those segments of nucleic acid sequences that can form these structures. To identify this, an algorithm is required that can find the possible quadruplex forming sequences (PQFSs). Though there are some algorithms for this purpose, such as Quadparser, Quadfinder and QuadBase, another algorithm is proposed that not only find the PQFSs but it has an additional feature to map the transcripts and their exons for PQFSs.

As a gene is composed of one or more exons, there is a possibility to find PQFS when two exons join each other after removal of introns (known as exon-exon junction). So, the QuadLocator, the proposed algorithm, introduces this additional feature that provides the transcript-exon mapping to check if there are PQFSs at exon-exon junctions. It is aimed to provide this tool as a web application in future.

REFERENCES:

Armitage B. A. (2007) *The rule of four*, Nature Chemical Biology, 3(4), 203-204.

Bates P., J. L. Mergny, D. Yang (2007) *Quartets in G-major*, EMBO reports, 8(11), 1003-1010.

Darnell J. C., K. B. Jensen, P. Jin, V. Brown, S. T. Warren, R. B. Darnell (2001) *Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function*, Cell, 107(4), 489-499.

Huppert J. L., S. Balasubramanian (2005) *Prevalence of quadruplexes in the human genome*, Nucleic Acids Research, 33(9), 2908–2916.

Karimata H., D. Miyoshi, N. Sugimoto (2005) *Structure and stability of DNA quadruplexes under molecular crowding conditions*, Nucleic Acids Symposium Series, Oxford University press, 49(1), 239-240.

Langdon W., G. Upton, A. Harrison (2009) *Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips*, Briefings in bioinformatics, 10, 259-277.

Memon F. N., O. Sanchez-Graillet, G. J. G. Upton, A. M. Owen, A. P. Harrison (2010a) *Identifying the impact of G-Quadruplexes on Affymetrix 3′ Arrays using Cloud Computing*, Journal of Integrative Bioinformatics , 7(2), 111Pp.

Memon F. N., G. J. G. Upton, A. P. Harrison (2010b) *A comparative study of the impact of G-stack probes on various Affymetrix GeneChips of mammalian*, Journal of Nucleic Acids special issue on G-Quadruplex Nucleic Acids, Volume 2010 (2010), Article ID 489736, 6 pages http://dx.doi.org/10.4061/2010/489736.

Neidle S., S. Balasubramanian (2006) *Quadruplex nucleic acids (Book)*. Royal Society of Chemistry, RSC Sciences Series, Volume 7 of RSC biomolecular sciences.

Qin Y., L.H. Hurley (2008) *Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions*, Biochimie, 90(8), 1149-1171.

Scarlata V., M. Hariharan, A. Arora, S. Maiti (2006) *Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences*, Nucleic Acids Research, 34(suppl 2), W683-W685.

Shanahan H., F. N. Memon, G. Upton, A. Harrison (2012) *Normalized Affymetrix expression data are biased by G-quadruplex formation*, Nucleic Acid Research, 40(8), 3307- 3315.

Siddiqui-Jain A., C. L. Grand, D. J. Bearss, L. H. Hurley (2002) *Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription*, Proceedings of the National Academy of Sciences of the United States of America, 99(18), 11593–11598.

Upton G., W. Langdon, A. Harrison (2008) *G-spots cause incorrect expression measurement in Affymetrix microarrays*, BMC Genomics, 9, 613, 10 pages, doi:10.1186/1471-2164-9-613.

Upton G. J. G., O. Sanchez-Graillet, J. Rowsell, J. M. Arteaga- Salas, N. S. Graham, M. A. Stalteri, F. N. Memon, S. T. May, A. P. Harrison (2009) *On the causes of outliers in Affymetrix GeneChip data*, Briefings in Functional Genomics and Proteomics, 8(3), 199-212.

Wu, C., H. Zhao, K. Baggerly, R. Carta, L. Zhang (2007) *Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays*, Bioinformatics, 23, 2566.

Yadav, V. K., J. K. Abraham, P. Mani, R. Kulshrestha, S. Chowdhury (2008) *QuadBase: genome-wide database of G4 DNA occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes*, Nucleic Acids Research, 36(Database issue), D381-D385.

Zahler A. M., J. R. Williamson, T. R. Cech, D. M. Prescott (1991) *Inhibition of telomerase by G-quartet DMA structures*, Nature, 350(6320), 718-720.