Detection of G-type density in promoter sequence of colon cancer oncogenes and tumor suppressor genes

Senol Dogan1*, Anis Cilic1, Amina Kurtovic-Kozaric1,2 & Fatih Ozturk3

1Department of Genetics and Bioengineering, International Burch University, Francuske revolucije BB, Ilidža 71000; 2Department of Clinical Pathology, Clinical Center of the University of Sarajevo, Bosnia and Herzegovina; 3Department of Information Technologies, International Burch University, Francuske revolucije BB, Ilidža 71000, Sarajevo, Bosnia and Herzegovina; Senol Dogan – Email: sdogan@ibu.edu.ba; Phone: 0038762801777; *Corresponding author

Received May 07, 2015; Revised May 13, 2015; Accepted June 16, 2015; Published June 30, 2015

Abstract:
The guanine rich locations are present in human genome. Previous studies have shown that the presence of G rich sequences and motifs may be significant for gene activity and function. We decided to focus our interest to identify G rich motifs in promoters of oncogenes and tumor suppressor genes. We used a set of 100 most common oncogenes and tumor suppressor genes (TSG) for this analysis. We collected 600nt long promoters with -500 and +100 TSS (transcription start site) from the oncogenes and TSG set. Using a computer program, we calculated the G densities using numbers and locations of G forms with 100nt moving widow. We included G numbers from 2 to 7 guanines. Analysis shows that G density increases from -500 to +100 and more from TSS. G density is found to be maximum within -/+100 of TSS. The results of G densities were compared with the expression data of the selected oncogenes and tumor suppressor genes in patients with colon cancer (n=174).

Background:
The guanine rich region is a relatively unexplored part of the human genome. Although there are some algorithms to detect special motifs, such as G quadruplex, the algorithms to detect other types of G rich motifs do not exist. It was first reported in 1910 that guanylic acid forms a gel at high concentrations [1]. Therefore, it is suggested that G-rich sequences may form some other structures. About 50 years later, Gellert used X-ray diffraction to display that guanylic acids can accumulate into tetrameric structures [1]. The presence of G-rich sequences is found in functional regions of many genomes. For example, G-rich regions have the potential to form G4 structures which locate telomeres, promoters, mitotic and meiotic double-strand break (DSB) sites [2]. Naturally occurring ‘G’ rich sequences, via non-Watson-Crick base pairing capable of forming G-quadruplexes and stabilized by cyclic Hoogsteen hydrogen bonding, have been implicated in some different genomic activities such as: transcription pausing, FMRP binding, mRNA stability, translation initiation as well as repression [3]. Although the G-quadruplex (G4) motif has been analyzed as a non-B-form DNA secondary structure, there may be some others which have not been given nomenclature yet [4]. It is already known that G-quadruplex is involved in different human cancers [5, 6, 7]. Thus G-quadruplexes may be targeted for therapeutic purposes [5, 6, 7]. DNA folding properties allow it to make various inter- and intramolecular secondary structures. Although the structures seem in vitro artefacts, bioinformatics reveals that DNA sequences capable of forming such structures which are conserved [2].

It is known that there are some types of guanine rich regions and motifs. Z-DNA motifs are mostly related to transcriptional start sites in eukaryotic genomes [8]. Cruciform structures are located close to replication origins, breakpoint junctions and promoters in various organisms. Triplexes cause genomic instability by breaking double-strand that result in translocations [9]. The repeated expansion may relate to human genetic disorders [10]. G4 structures present different...
topologies and are separated into various groups depending on the orientation of the DNA sequences. It is unclear how many G-rich sequences form stable G4 structures in vivo, but G4 DNA motifs are common in G-rich micro and minisatellites, up and downstream of TSSs, often near promoters, transcription factor binding sites, and mitotic and meiotic DSB sites [11, 12, 13, 14]. Telomerase activity in most human cancers can be influenced by G4, because different small ligands target the regions and bind, as has been tested in different experiments [15]. G4 motifs are most likely found within 1,000 nt upstream of the TSS in 50% of human genes [16]. Special Bioinformatics algorithms find that the promoters of human oncogenes and regulatory genes have G4 motifs more than in the promoters of housekeeping and tumor suppressor genes [14]. G-rich sequences or G4 may cause supercoiling in the structures are in or near promoter regions, which can have both positive and negative effects on transcription. First, the location of the G motif is a very powerful factor for transcription.

Approximately ~ 400,000 presumed G4 motifs are found in the human genome. The motifs are frequently located within the promoter regions of oncogenes, assuming that G4 motifs may act in a key role for regulation of different cellular activities such as transcription, translation, telomere maintenance, and replication [2]. The G4 motif importance in the regulation of gene transcription came from v-myc viral studies, oncogene homolog (MYC), the transcription factor regulates the expression of different genes which are altered in human cancer, is a non-regulated in around half of the tumors [17]. Guanine-rich nucleic acid sequences which form G-quadruplex structures are key regulators of some biological processes and are targeted for therapeutic medicine such as Quarfloxin, a fluoroquinolone [18, 19].

Guanine numbers and densities are very distinctive parts of the genome. The aim of this study is to find repeating G motifs consisting of 2, 3, 4, 5, 6, and 7 guanines in the promoter sequence of selected genes important for carcinogenesis (50 tumor suppressor genes and oncogenes). Previous studies have analyzed G quadruplexed for oncogenes, but not other types of motifs and genes like tumor suppressor genes [20]. However, in this paper the promoter sequences of oncogenes and tumor suppressor genes (TSG) are the candidates for finding G-type densities.

**Methodology:**

**Data types**

Three different databases have been used, including Genecards, EPD (Eukaryotic Promoter Database), and TCGA (The Cancer Genome Atlas). The workflow has been described as a flow chart (Figure 1). The names of the oncogenes and tumor suppressor genes (TSG) most related to colon cancer are taken from Genecards [21]. Fifty oncogenes and fifty TSGs are
selected from Genecards for each group. According to the chosen oncogenes and TSGs, promoter sequences are downloaded from EPD [22]. The genes’ promoter sequences consist of 600 nucleotides, -500 before Transcription Start Site, (TSS), and +100 after the Transcription Start Site. The database, EPD, which promoter sequences are downloaded has supplied the promoter sequence 500 before TSS and 100 after TSS [22]. The cancer genomic data portal is TCGA [23] from which 174 colon cancer patients with gene expression Level 3 and the 46 control data are downloaded on 01/02/2015.

Figure 2: 100 window promoter sequences. The G types (G2,7) number is detected in both Oncogenes (A) and TSG (B). The number shows conservative structure, almost together high and low. Panel A represents the oncogenes G types and number. The number increases as closer to TSS and G2 is getting the maximum number. Panel B represents the TSG G types and number. Almost all types dramatically increase between 200-100 to 100-0 windows.

**Guanine density detection**

The guanine nucleotide number is reported in other studies, especially in genomic locations such as, telomere, promoter, exon and intron [24]. G types including GG, GGG, GGGG, GGGGG, GGGGGG, GGGGGGG, have been produced to detect guanine density (GD) in the promoter sequences [25]. For each promoter sequence, GD is detected by a computational program which was created for this study. The program searches GD types of the sequences between -500 to +100 in a 100 nucleotide group for both oncogenes and TSG in Figure 2. According to the guanine density of each group, oncogenes and TSG promoter sequence profiles are characterized and listed Table 1 (see supplementary material).

Our results indicate that the oncogenes and TSG G profiles present increasingly high density between -100 to 0, where they achieved maximum density after Transcription Start Site (TSS) (Table 1 & Table 2, respectively). The G types, G2, G3, G4, G5, G6, G7, show increasing order and reach the maximum level after TSS, 0 to +100. The G-types show diverse density in different groups of the promoter sequences. Especially, G2, G3, G4, G5 types are detected more than other types. In addition to that, in the -100 to 0 locations of the sequence, the G6 types appear 8 times in both oncogenes and TSG, which is especially rare. G2 is the most commonly found type and followed G3, G4 and G5 in the all small groups. Unexpectedly, G7 type is found 2 times between -200 to -100 in both promoter sequences (Table 1 & Table 2 (see supplementary material)).
If we analyze the promoters of oncogenes only, the G profiles consist of maximum G3, G5, G6, G7 types in -100-0 and G2, G4 types maximum 0-100. In the TSG promoters, the maximum profile of G types is demonstrated as G4, G5, G6, G7 in -100-0 and G2, G3 in 0-100. The maximum G density is found before and after 100 nucleotides after TSS (Figure 3).

G-type density is compared before and after TSS; the average G-type density of all 5 nucleotide groups between -500 and 0, is compared to the G-type density of the group between 0 and +100. Before TSS, on average, starting from -500, the oncogenes have 336 and TSG have 332 G types, but after TSS, to +100, the oncogenes have 402 and TSG have 435 G types (Figure 4). Surprisingly, 200-100 location of both promoters has distinct G types, such as G9 and G11 in oncogenes and 2 times G8 in TSG. The G types number increases between the segments 400-300 to 300-200 around and then a little bit decrease after that. However, the last segment before the TSS and 100 nucleotides after TSS have the maximum G type number over all comparison.

**Gene expression comparison**

Since G profiles are found in promoters sites, which are important for the regulation of gene expression, we decided to compare the G profiles of selected cancer-related genes to their expression in the colon cancer patients. The expression data were downloaded from The Cancer Genome Atlas (TCGA) data portal which supplies many different patient genomics data, including gene expression, microRNA, RNAseq, methylation, mutation and others. We downloaded the expression data from 17815 genes from 174 colon cancer patients and 46 controls. The average expression level of all 17815 genes was determined and compared with control data (Supplementary 2). Abnormal fold change of 50 selected oncogenes and 50 selected TSGs has been found. According to the fold change, high-expressed and low-expressed genes are profiled with G Density Table 3 (see supplementary material).

**Discussion:**

Guanine numbers and densities are very distinctive parts of the genome. In this study, we presented the G densities of motifs consisting of 2, 3, 4, 5, 6, and 7 guanines in the promoter sequence of selected genes important for carcinogenesis (50 tumor suppressor genes and 50 oncogenes). Previous studies presented different algorithms and methods to find guanine rich regions and potential motifs. In those studies, different G-score for G quadruplex calculation methods were developed [20]. However, no previous study has shown the densities of other G repeats such as GG, GGG, GGGG, GGGGG, GGGGGG, GGGGGGGG and GGGGGGGGG. Our study is the first to compare the G repeats in tumor suppressors and oncogenes’ promoters.

The promoter sequences were separated into small groups of 100 nucleotides, from -500 to +100. Our results showed that the oncogenes and TSG G profiles present increasingly high density between -100 to 0, where they achieved maximum density after Transcription Start Site (TSS) (Table 1 & Table 2, respectively). Analysis shows that G density increases from -500 to +100 and more from TSS. G density is found to be maximum within -/+100 of TSS. The results of G densities were compared with the expression data of the selected oncogenes and tumor suppressor genes in patients with colon cancer (n=174, Table 3).

**The relation between gene expression and Guanine types density**

Since G profiles are found in promoters sites, we decided to compare the G profiles of selected cancer-related genes to their expression in the colon cancer patients. TCGA colon cancer data of 174 patients and 46 controls are compared. According to fold changes, high and low expressed genes have been determined as compared to the controls. All types of G repeats of 18 highly expressed genes from both oncogenes and TSG have been analyzed Table 3 (see supplementary material). The 18 highly expressed oncogenes have the average of all G-types to be 44.44 and the low expressed to be 46.17. On the other hand, the 18 highly expressed TSGs have the average of all G types to be 41.17 and the low expressed to be 46.94.

**Conclusions:**

This study describes a method with a computer program for quantitatively evaluating the conservation of different guanine...
types and densities. Guanine types, G (2-7), can be identified by guanine density (GD) program in order to detect potential sequence motifs which are conserved in promoters of oncogenes and TSGs. The computer program quickly and efficiently identifies conserved Guanine Types Density regions where there is a relatively high probability of sequence conservation. The program reported in this study has application for the analysis of large datasets.

Our results show that depending on the exact locations of the guanine types and density, the gene promoter sequences demonstrate conserved characteristics. In other words, the G densities increase closer to the transcriptional start site of both oncogenes and TSGs. The G density is the highest within the 100 base pairs proximal to the transcriptional start site. Identifying common conserved GDs may help us validate these findings on larger datasets to show the role of G densities in pathogenesis and disease.

Moreover, the G types density demonstrates that the location and number of G repeats are conserved in oncogenes and TSG promoter sequence. The paper may help elucidate the potential role of the specific G types in therapeutic and diagnostic pursuits.

Reference:
[1] Gellert M et al. Proc Natl Acad Sci. 1962 48: 2013 [PMID: 13947099]
[2] Bochman ML et al. Nat Rev Genet. 2012 13: 770 [PMID: 23032257]
[3] Davis JT et al. Angew Chem Int Ed Engl. 2004 30: 668 [PMID: 14755695]
[4] Svozil D et al. Nucleic Acids Res. 2008 36: 3690 [PMID: 18477633]
[5] Bagga JS et al. Hum Genomics. 2013 7: 19 [PMID: 24040966]
[6] Wu Y et al. FEBS J. 2010 277: 3470 [PMID: 20670277]
[7] Balasubramanian S et al. Curr Opin Chem Biol. 2009 13: 345 [PMID: 19515602]
[8] Schroth GP et al. J Biol Chem. 1992 267: 11846 [PMID: 1601856]
[9] Wang X & Haber JE, PLoS Biol. 2004 2: 21 [PMID: 14737196]
[10] Rolfsmeier ML et al. Mol Cell. 2000 6: 1501 [PMID: 11163222]
[11] Capra JA et al. PLoS Comput Biol. 2010 6: e1000861 [PMID: 20676380]
[12] Hershman SG, Nucleic Acids Res. 2008 36: 144 [PMID: 17999996]
[13] Nakken S et al. Nucleic Acids Res. 2009 37: 5749 [PMID: 19617376]
[14] Eddy J & Maizels N, Nucleic Acids Res. 2006 34: 3887 [PMID: 16914419]
[15] Neidle S et al. FEBS J. 2010 277: 1118 [PMID: 19951354]
[16] Huppert JL & Balasubramanian S, Nucleic Acids Res. 2007 35: 406 [PMID: 17169996]
[17] Delgado MD et al. Clin Transl Oncol. 2013 15: 87 [PMID: 22911553]
[18] Collie GW & Parkinson GN, Chem Soc Rev. 2011 40: 5867 [PMID: 21789296]
[19] Bugaut A & Balasubramanian S, Nucleic Acids Res. 2012 40: 4727 [PMID: 22351747]
[20] Frees S et al. Hum Genomics. 2014 8: 10 [PMID: 24885782]
[21] http://www.genecards.org
[22] http://epd.vital-it.ch
[23] https://tcga-data.nci.nih.gov/tcga/
[24] Burge S et al. Nucleic Acids Res. 2006 34: 5402 [PMID: 17012276]
[25] Bidzinska J et al. Molecules. 2013 18: 12368 [PMID: 24108400]

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License.

Edited by P Kangueane

Citation: Dogan et al. Bioinformation 11(6): 290-295 (2015)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 11(6):290-295 (2015) 294 © 2015 Biomedical Informatics
### Table 1: G Density in Promotor Sequences of Oncogenes

| Oncogenes | Guanine Density | Range of The Promoter Sequence from -499 to 100 bp relative to TSS |
|-----------|----------------|---------------------------------------------------------------|
|           | -500 to -400   | -400 to -300 | -300 to -200 | -200 to -100 | -100 to 0 | 0 to 10 |
| GGGGGGG   | 0              | 0           | 0           | 2           | 2         | 0       |
| GGGGGG    | 0              | 1           | 1           | 1           | 8         | 1       |
| GGGGG     | 5              | 11          | 15          | 15          | 16        | 4       |
| GGGG      | 35             | 33          | 45          | 41          | 27        | 37      |
| GGG       | 68             | 70          | 85          | 85          | 75        | 69      |
| GG        | 204            | 203         | 202         | 201         | 230       | 291     |

### Table 2: G Density in Promotor Sequences of TSG

| TSG | Guanine Density | Range of The Promoter Sequence from -499 to 100 bp relative to TSS |
|-----|----------------|---------------------------------------------------------------|
|     | -500 to -400   | -400 to -300 | -300 to -200 | -200 to -100 | -100 to 0 | 0 to 10 |
| GGGGGGG | 0              | 0           | 0           | 2           | 1         | 0       |
| GGGGGG  | 0              | 3           | 1           | 1           | 8         | 1       |
| GGGG    | 8              | 9           | 9           | 5           | 14        | 6       |
| GGG     | 33             | 28          | 39          | 24          | 41        | 36      |
| GG      | 56             | 77          | 76          | 84          | 89        | 92      |
| GG      | 214            | 195         | 194         | 196         | 254       | 300     |

### Table 3: High and Low Expression of Oncogenes and TSG with G Types Density

| Oncogenes (expression) | TSG (expression) | High | Low |
|------------------------|------------------|------|-----|
|                        |                  |      |     |
| CDKN1A                 | TGFBR1           | 68   | 75  |
| BMI1                   | ODC1             | 65   | 62  |
| MET                    | MYB              | 62   | 61  |
| BRAF                   | KRAS             | 52   | 57  |
| PI3M3                  | CTNNB1           | 51   | 53  |
| SNCG                   | GRP              | 49   | 52  |
| PI1M1                  | PTGS2            | 49   | 43  |
| ETV4                   | BIRC5            | 42   | 42  |
| EGFIR                  | TFF3             | 42   | 42  |
| MLH1                   | SSTR2            | 41   | 42  |
| PI2M2                  | CEACAM5          | 39   | 41  |
| TFF1                   | MYC              | 37   | 39  |
| NTRK1                  | CCND1            | 36   | 38  |
| TMY5                   | CDH1             | 36   | 37  |
| RET                    | CEACAM3          | 34   | 37  |
| MYCN                   | CDX2             | 33   | 37  |
| CDKN2A                 | SSTR1            | 32   | 37  |
| NTRK2                  | ERBB2            | 32   | 36  |
|                        |                  |      |     |
| Average                |                  | 44.44| 46.17|
| Average                |                  | 41.17| 46.94|

© 2015 Biomedical Informatics