Published in final edited form as:

Methods Mol Biol. 2010 ; 608: 39–50. doi:10.1007/978-1-59745-363-9_3.

Computational Approaches to the Detection and Analysis of Sequences with Intramolecular G-Quadruplex Forming Potential

Paul Ryvkin, Steve G. Hershman, Li-San Wang, and F. Brad Johnson

Abstract

Sequences with the potential to form intramolecular G-quadruplexes (G4-structures) are found in highly nonrandom distributions in the genomes of diverse organisms. These sequences are associated with nucleic acid metabolic processes ranging from transcription and translation to recombination and telomere function. Here we review different computational methods for identifying potential G4-forming sequences and provide protocols for their implementation. We also discuss methods for assessing the significance and specificity of associations between the sequences and different biological functions.

Keywords

G-quadruplex; G4-DNA; Bioinformatics; Computational biology

1. Introduction

G4-structures, including G4-DNA and G4-RNA, comprise stacked quartets of Hoogsteen hydrogen bonded guanines stabilized by small monovalent cations. Interest in G4-structures has been sparked by recent findings that suggest that they function in processes ranging from transcription and translation to recombination and telomere maintenance (1, 2). A large number of particular G4-structures are possible, even given the same starting sequences (3, 4). For example, G4-structures can be assembled from guanines within one nucleic acid strand (intramolecular) or from different strands (intermolecular). They can also differ on the basis of the glycosidic bond angles of the guanines, the type of coordinating cations, the number of stacked quartets, the polarity of phosphodiester backbone strands, and the length, nucleotide composition, and arrangement of the sequences that do not contribute to the quartets. This structural diversity has made it challenging to predict which sequences are most likely to form stable G4-structures, although studies of oligonucleotides in vitro have generated some simple rules, e.g. intramolecular G4-DNA is favored when one or more of the intervening loops are short (5). Furthermore, G-quadruplex formation in vivo is likely to be modulated by chromatin factors, including proteins and helicases that can stabilize or unwind G4-structures (6). Therefore, it is currently not possible to predict with certainty whether a given sequence will form G4-structures within a living cell.

Computational approaches can nonetheless be of value for indicating the potential of a sequence to form G4-structures. There are some sequences that have essentially no potential for forming G4-structures, and computational approaches can be used to separate these from other sequences with higher G4-forming potential. The question can then be asked whether the two sets of sequences are differentially associated with any biological activities that might be of interest, e.g. expression under certain circumstances, chromatin modifications, recombination hotspots, polarity of replication, conservation among or within species, etc. Although such an association will not prove that it is the capacity of the sequences to form G4-structure per se that explains the association, it does provide a starting point for hypothesis

testing. Moreover, if the biology being explored involves factors that a priori are expected to influence G4-structure levels or function, an association with sequences having high G4-forming potential provides an argument for the importance of the G4-structures themselves. This interpretation can be strengthened if the associations with sequences having G4-forming potential are stronger than with carefully selected control sequences.

Computational approaches for identifying sequences with G4-forming potential are currently based primarily on three simple rules derived from studies of intramolecular G4-DNA formation in vitro (7–11). (1) At least two and preferably three or more G-quartets must be stacked to form a stable G-quadruplex (3); for intramolecular G4-DNA this translates to runs of two-to-three or more consecutive Gs separated by three loop sequences. (2) Short loops (one to two nucleotides (nt)) favor G4-DNA formation, although longer loops are possible (5). Indeed, at least when one or two of the loops are short, the other(s) can extend beyond 20 nt (12). (3) To a first approximation, the loop sequences are not critical determinants of stability, although recent investigations are beginning to define their contributions (13). Here we describe several of these computational approaches and how to calculate the statistical associations of particular biological variables with the identified potential G4-forming sequences, and discuss controls that can help evaluate the specificity of any observed associations. Because of the exploratory nature of these approaches, the methods presented should not be considered definitive, but it is hoped that they will stimulate collaborative efforts among molecular and computational biologists. In addition, we emphasize that in many cases these approaches should be considered only a starting point for generation of hypotheses that can then be tested using molecular genetic and biochemical approaches.

2. Methods

2.1. Detecting Potential Intramolecular G-quadruplex-Forming Sequences

Three types of computational approaches, based on string pattern matching, have been used in the literature for analyzing the genomic distribution of sequences with the potential to form G-quadruplexes. Although G4-structures could form between different nucleic acid strands, e.g. conceivably between the two strands of a denatured DNA duplex, we limit our discussion to sequences with the potential to form intramolecular G-quadruplexes because all the algorithms published to date are for identifying intramolecular structures.

2.1.1. Regular Expression—This method specifies a regular expression that the G-quadruplex-forming sequence should take. For example, the first analyses of G-quadruplex forming potential of the human genome (8, 9) used the regular expression $G_{3+N_{1-7}} G_{3+N_{1-7}} G_{3+N_{1-7}} G_{3+}$, which requires a matching sequence to satisfy two properties: (1) each of the four guanine runs has a length of at least three nucleotides, and (2) lengths of the three loops are all between 1 and 7 nucleotides (N means any nucleotide). Many programming languages provide regular expression matching; the following example of Perl code:

```
# search this sequence
$s = "AATACGGGACATGGGGATAGAGGCGCGGGTT";
if ($s =~ m/G{3,}.{1,7}?G{3,}.{1,7}?G{3,}.{1,7}?G{3,}/){print "Match";}
}
```

Detecting G4-forming candidate sequences on the complementary strand can be done easily by replacing Gs with Cs in the regular expression above. The output of a regular expression analysis is discrete, with any particular stretch of DNA either conforming to the pattern or not. However, once individual potential G4-forming sequences have been identified, their density

within any region (e.g. a promoter or intron) can be assessed to provide a more continuous estimate of the overall G4-forming potential of the region. For example, the fraction of the region contained within sequences having G4-forming potential can be calculated. Alternatively, the number of such sequences in a region can be divided by the length of the region; with this approach an arbitrary decision needs to be made for whether overlapping sequences will be counted as individual or distinct occurrences.

2.1.2. G4P—Eddy and Maizels (10) used a sliding window approach to assess the G-quadruplex-forming potential of genomic regions. The algorithm generates a continuous estimate of G4-forming potential that depends on three parameters, k (length of the guanine runs; default is 3 nt), w (window size; default is 100 nt), and s (step; default is 20 nt). Starting from the beginning of the input sequence, the algorithm checks windows of length w starting every s nucleotides; the G4P is the fraction of these windows containing four guanine runs of length k separated by at least one nucleotide. This approach is more flexible than the regular expression, because it only limits the total length of the candidate sequence and not the lengths of individual loops. The authors have made public a program for Microsoft Windows (<http://depts.washington.edu/maizels9/G4calc.php>); the program calculates G4P on either strand, or the average of the G4P on both strands. The following example shows how G4P (and C4P) can alternatively be computed using the `gregexpr` command in R.

```
ins = paste(sample(c("A", "C", "G", "T"),
1000000, prob = c(0.3, 0.2, 0.2, 0.3), repl = T),
collapse = "") # random sequences
gpat = "G{3}.*?G{3}.*?G{3}.*?G{3}";
cpat = "C{3}.*?C{3}.*?C{3}.*?C{3}";
n = nchar(ins);
k = n/20
fwdcnt = rep(0, k + 1); revcnt = rep(0, k + 1); #
window match results
for (i in 0:k) {
  ins.k = substring(ins, i*20 + 1, min(i*20 + 100,
n)) # window i
  if (gregexpr(gpat, ins.k, perl = T)[[1]][1]
!= -1) {fwdcnt[i + 1] = 1}
  if (gregexpr(cpat, ins.k, perl = T)[[1]][1]
!= -1) {revcnt[i + 1] = 1}
}
g4p = sum(fwdcnt)/length(fwdcnt)
c4p = sum(revcnt)/length(revcnt)
```

2.1.3. QFP—Recently we described the distribution of sequences with G-quadruplex forming potential (QFP) within the *Saccharomyces cerevisiae* genome (11). We used the sliding window approach, but instead of returning a continuous estimate of G4-forming potential, we identified the discrete sites that may form G-quadruplexes. In other words, for $k = 3$ and $w = 100$, the sequence takes the form $G_3 N_a G_3 N_b G_3 N_c G_3$ where $a, b, c > 0$, and $12 + a + b + c = 100$. This approach is more sensitive than the regular expression and G4P approaches, which are useful for an organism like yeast that generally has rare sequences with G4-forming potential. An efficient algorithm examines every sliding window that starts from a GGG run, instead of running sliding windows starting at every nucleotide position. Software is available from hershman@fas.harvard.edu.

Comparisons of the three methods are shown in Fig. 3.1 and Table 3.1.

2.2. Analysis of the Genomic Distribution of G-quadruplex-Forming Sequences

Once G4-forming candidate sequences are identified or G4 forming potentials (G4Ps) are determined, it may be of interest to compare their distributions with other genomic regions, e.g. promoters, and, moreover, with the behavior of these genomic regions in particular biologic settings. For example, if genes that are differentially expressed under certain experimental conditions have a statistically significant tendency to have G-quadruplex forming sequences in their promoter regions, this raises the possibility that the G4-forming sequences play some role in the regulation of expression. Furthermore, if the experimental conditions being investigated are those that are expected a priori to influence G4 levels, e.g. treatment of cells with small molecule G4 ligands or deletion of a G4-DNA unwinding helicase, then a statistical association between loci with altered expression and loci with potential G4-forming sequences indicates that G-quadruplex formation by the sequences may be part of the mechanism of differential gene expression. Care should be taken when computing the significance of associations. Depending on the types of G-quadruplex detection (discrete or continuous potential), different statistical tests must be used. Also, we note that although we are using transcription as the biological variable in these examples, similar approaches can be applied to other processes (e.g. polarity of replication, recombination rates, etc).

2.2.1. Continuous Variables—If the association to be tested involves continuous *t*-test statistics for gene expression and a continuous measure of G4-forming potential (i.e. G4P or a continuous derivative of output from a regular expression), one can run a test of correlation using Pearson's correlation (assuming normally distributed error) or Spearman's correlation (nonparametric test without assumption of normality). The following is an example in R code:

```
# vector tstat stores the t-statistics for
all genes in a
# control/treatment microarray experiment
# vector g4p stores the G4P scores for all
genes
p.cor = cor(tstat,g4p) #Pearson
correlation
p.pval = cor.test(tstat,g4p)$p.value
#Significance
s.cor = cor(tstat,g4p,method = "spearman")
#Spearman
correlation
s.pval = cor.test(tstat,g4p,
method = "spearman")$p.value
#Significance
```

2.2.2. Discrete Variables—At the other end of the spectrum, associations can be tested between discrete categories for altered gene expression and for G4-forming potential, e.g. QFP or discrete output from a regular expression. In such cases each gene will fall into one of four categories based on whether the gene is differentially expressed and whether the gene has a G-quadruplex-forming sequence. A 2-by-2 contingency table can be formed for the number of genes falling into the four combinations, and then either a one-sided Fisher's test (if there is an expectation for the direction of the association, e.g., G-quadruplexes activate gene expression) or a two-sided Fisher's test (if no expectation regarding the direction of association; in which case the odds ratio returned from Fisher's test shows the direction of association) can be applied. Many computer programs, such as R (the `fisher.test` function), provide this

capability: # vector diff_expressed is a 0/1 vector for the status of expression # change (1 if differentially expressed) for all genes in a # control/treatment microarray experiment # vector g4p stores the G4P scores for all genes tbl = data.frame(diff_expressed, g4p) # 2-column table, one row per gene r = glm(diff_expressed~g4p, data = tbl,family= "binomial") # logistic regression summary(r) # summary of regression; check the p-value # for the coefficient of g4p for # association significance

2.2.3. Mixed Continuous and Discrete Variables—A middle ground also exists, involving mixtures of continuous and discrete categories for G4-forming potential and altered expression. For example, the association of G4P with whether each gene is differentially expressed or not can be tested. One might be tempted to divide the G4Ps into two groups and run a *t*-test, but statistically this is not correct as genes in the two groups are not independent observations from two populations. A better approach is to run a logistic regression using the G4P as the independent variable and the differential expression status as the dependent variable. We can test similarly if we test the association using a discrete G4 criterion (QFP or regular expression) and a continuous score for differential gene expression such as the *t*-statistic:

```
# vector has_gquad is a 0/1 vector on whether
each gene has a
# G-quadruplex motif (1 if the gene has a
G-quadruplex motif)
tbl = data.frame(diff_expressed, has_gquad) #
two-column table,# one row per gene
pval = fisher.test(tbl)$p.value # Significance
pval2 = fisher.test(tbl, alt="greater")$p.value
# Significance, one-sided
```

Usually 0.01 or 0.05 is used as a threshold for significance. When performing multiple tests, the threshold of significance can be set using the Bonferroni correction, which involves simply dividing the threshold of individual tests, e.g., 0.05, by the number of tests to give a lower threshold. Alternatively, more sophisticated multiple testing correction procedures such as false discovery rate (FDR) can be used (14); the *fdrtool* package for R (15) includes an implementation.

Note that each statistical test returns the test statistic in addition to the statistical significance. While the significance shows whether the association arises by chance, the interpretation of the statistic is equally important. For example, the Pearson's correlation coefficient test returns the *p*-value indicating that the actual correlation coefficient is different from zero, and a very weak correlation may be statistically significant when many observations are made, which is often the case in genome-scale analyses such as microarray experiments. Thus biological and statistical significance are separate entities that must be interpreted carefully.

2.3. Controls to Assess the Specificity of Associations with G4-Forming Sequences

In addition to the association significance introduced in the last section, a common approach to assess the specificity of associations between potential G4-forming sequences and particular genomic features or biological functions is to make similar comparisons with control sequence patterns. The premise of this approach is that if G4-forming sequences are the sole (or primary) mediators of the biological function then there should be no (or less significant) association with control sequences. This premise will not always be true, e.g. the binding of transcription factors as well as G4-DNA formation might regulate transcription from a given promoter, and thus these control approaches can suffer from being too stringent. However, they can still be very informative, and we describe several commonly used controls below. We note that it is

important that the frequency of control patterns is on the same order as the frequency of the G4 pattern; otherwise, the more frequent pattern will have the potential to show greater statistical significance only on the basis of a larger number of observations. We will describe several commonly used controls as follows.

2.3.1. A4/T4 Control—One seemingly simple control is to examine the distribution of sequences having the form of the G4-forming potential strings, but with the Gs replaced by As or Ts, e.g. in the case of the regular expression the sequence $A_{3+}N_{1-7} A_{3+}N_{1-7} A_{3+}N_{1-7} A_{3+}$ (8). Assuming there is no biological significance for such “A4” or “T4” sequences, these control sequences should not be associated with the biological variables under investigation. In practice this approach may be problematic for two reasons: (a) in A/T-rich genomes the number of control sequences will be very high, e.g. A4/T4 sequences are nearly an order of magnitude more frequent in the human genome than G4/C4 sequences and this may inflate the significance of even very weak associations, and (b) it is possible that A4 or T4 sequences will have some functional connection to the biology being investigated.

2.3.2. Randomized Models—A second control approach is to generate random sequences that share characteristics of real genomes (8, 11). This approach is most useful for assessing whether the observed distribution of potential G4-forming sequences is nonrandom, but may also have some applications in association studies. For example, G4-forming sequences within randomized model genomes should be less associated with the biology under investigation than similar sequences in the real genome if the G4-forming potential and not some other related feature of the sequences (e.g. GC-richness) is responsible for the association. There are several possible approaches to generating randomized genomes. For example, a higher-order Markov chain based on real genomic sequence can be used to generate random sequences that share the same base composition, and diad, triad, etc., frequencies as the real genome. There are many software programs that can accomplish this (16), and one can implement first or second order Markov chains easily using statistical software such as R. This control is limited by the level of realism achieved by the Markov model and is more complex to implement than the others we describe. We therefore suggest researchers with limited experience in programming to seek collaboration with a statistician or a computer scientist for this control.

2.3.3. Partial G4 Controls—A third approach is to make comparisons to a sequence motif that is very close to a potential G4-forming sequence but which should not be able to form G4-structures. We suggest two methods. One is to examine strings with three G_{3+} runs instead of four. For example, in the case of the regular expression $G_{3+}N_{1-7} G_{3+}N_{1-7} G_{3+}N_{1-7} G_{3+}$, the pattern $G_{3+}N_{1-7} G_{3+}N_{1-7} G_{3+}$ could be examined with the added requirement that there be no G_{3+} run within eight nucleotides of the pattern. A second approach is to examine strings that have G-runs interrupted by non-G bases. The idea is to find sequences that have an edit distance of n (point mutations) from putative quadruplex-forming sequences, where $n = 1$ or 2 is ideal. The following Perl code can perform such a search:

```
$s = "AATACGGGACATGAGGATAGAGGGCGCGGGTT" ;
# search string
# interrupted G-runs
if ($s =~ /(G.G){1,7}(G.G){1,7}(G.G){1,7}
(G.G)/) {
$runs = "$1$2$3$4"; # concatenate runs
$dist = () = $runs =~ /[^\G]/g; # count non-G's in
runs
if ($dist == 1) { print "Match"; }
}
```


For association studies it is important to consider that loci can have both potential G4-forming sequences and the control sequences, and therefore loci having the control sequence but not the G4-sequence must be identified to provide the control set.

2.3.4. Relaxed G4-Potential—A fourth approach is to relax the stringency for detection of potential G4-forming sequences, to see if the statistical significance dissipates, as the stability of intramolecular G4-structures should depend on how close the GGG runs are to one another. For example the loop lengths can be increased beyond a regular expression or the size of the sliding window can be increased in G4P or QFP detection. We used this method in our paper using QFP, and observed that the significance increases when the window size increases from 35 nt to 100 nt, but the significance drops afterwards (11). The initial positive correlation may be due to the low frequency of QFP sequences in the *S. cerevisiae* genome, and the loss of significance at windows larger than 100 nt due to more random targets (i.e. those not actually capable of forming G4 structures) being included.

2.3.5. Transcription Factor Comparisons—In cases involving associations between potential G4-forming sequences and gene expression an alternative explanation for the associations is that the G4 patterns simply reflect clusters of binding sites for transcription factors that bind to sequences containing G-runs. Similarly, the Gs in CpG islands can flank G runs and thus contribute to G4-forming potential sequences. Recently, Eddy and Maizels tested this idea by subtracting loci containing such patterns from their analyses and showed that this virtually eliminated the apparent associations with potential G4-forming patterns (17). However, it is an expected result that removal of sequences having the same pattern as those with G4-forming potential will diminish the significance of G4 associations. Furthermore, the correspondence between the presence of a potential binding site for a transcription factor and actual binding of the transcription factor is imperfect. Thus we argue that such approaches do not distinguish between the possibilities that the binding of transcription factors or G4-structure formation explains the associations. Data based on actual transcription factor occupancy (i.e. using chromatin immunoprecipitation), which are available on a large scale for model organisms like yeast (18), could help resolve this issue. Genetic approaches that manipulate the activity of such transcription factors could also provide valuable insights. More work is needed before the proper application of this control can be realized.

Determining which of the control approaches described above are most appropriate will require additional research. We suggest that the readers try several controls to better gauge the significance and specificity of their findings.

3. Resources on the Web

The G4P criterion was first defined in (10), and the program (Microsoft Windows only) can be downloaded from (<http://depts.washington.edu/maizels9/G4calc.php>). A regular expression program by Huppert et al. (8), *Quadparser*, can be obtained here: <http://www.quadruplex.org/?view=quadparser>. There is also a flexible web tool called *QGRS Mapper* (19) available at <http://bioinformatics.ramapo.edu/QGRS/analyze.php>. G4P and other methods in this chapter can be entirely implemented using either Perl (<http://www.perl.org>) or R (<http://www.r-project.org>). Example codes in this chapter can be downloaded from the companion website at http://people.pcbi.upenn.edu/~lswang/gquad/chapter_ryvkin_etal/.

Acknowledgments

We thank Jay Johnson, Kajia Cao, Marina Kozak, Alex Chavez, Jasmine Smith, and Qijun Chen for advice and discussions. This work was supported by NIH grants R01-AG021521, P01-AG031862, and a U. Penn Institute on Aging Pilot Grant.

References

1. Maizels N. Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat Struct Mol Biol.* 2006; 13:1055–1059. [PubMed: 17146462]
2. Johnson JE, Smith JS, Kozak ML, Johnson FB. In vivo veritas: using yeast to probe the biological functions of G-quadruplexes. *Biochimie.* 2008; 90:1250–1263. [PubMed: 18331848]
3. Lane AN, Chaires JB, Gray RD, Trent JO. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.* 2008; 36:5482–5515. [PubMed: 18718931]
4. Webba da Silva M. Geometric formalism for DNA quadruplex folding. *Chemistry.* 2007; 13:9738–9745. [PubMed: 17972263]
5. Bugaut A, Balasubramanian S. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry.* 2008; 47:689–697. [PubMed: 18092816]
6. Fry M. Tetraplex DNA and its interacting proteins. *Front Biosci.* 2007; 12:4336–4351. [PubMed: 17485378]
7. Rawal P, Kummarasetti VB, Ravindran J, Kumar N, Halder K, Sharma R, et al. Genome-wide prediction of G4 DNA as regulatory motifs: role in *Escherichia coli* global regulation. *Genome Res.* 2006; 16:644–655. [PubMed: 16651665]
8. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* 2005; 33:2908–2916. [PubMed: 15914667]
9. Todd AK, Johnston M, Neidle S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* 2005; 33:2901–2907. [PubMed: 15914666]
10. Eddy J, Maizels N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* 2006; 34:3887–3896. [PubMed: 16914419]
11. Hershman SG, Chen Q, Lee JY, Kozak ML, Yue P, Wang L-S, et al. Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 2008; 36:144–156. [PubMed: 17999996]
12. Bates P, Mergny JL, Yang D. Quartets in G-major. The First International Meeting on Quadruplex DNA. *EMBO Rep.* 2007; 8:1003–1010. [PubMed: 17901879]
13. Guedin A, De Cian A, Gros J, Lacroix L, Mergny JL. Sequence effects in single-base loops for quadruplexes. *Biochimie.* 2008; 90:686–696. [PubMed: 18294461]
14. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.* 2003; 100:9440–9445. [PubMed: 12883005]
15. Strimmer K. A unified approach to false discovery rate estimation. *BMC Bioinformatics.* 2008; 9:303. [PubMed: 18613966]
16. Ponty Y, Termier M, Denise A. GenRGenS: software for generating random genomic sequences and structures. *Bioinformatics.* 2006; 22:1534–1535. [PubMed: 16574695]
17. Eddy J, Maizels N. Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.* 2008; 36:1321–1333. [PubMed: 18187510]
18. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature.* 2004; 431:91–104.
19. Kikin O, D'Antonio L. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* 2006; 34:W676–W682. [PubMed: 16845096]

Fig. 3.1.
Example using three pattern matching criteria for intramolecular G-quadruplex candidate sequence detection.

Table 3.1

Comparison of three intramolecular G-quadruplex detection criteria

Criterion	Output format	Example criterion	Note
<i>Regular expression</i>	Matched sequence motif	G-run length 3, loop length 7	Most stringent of the three; imposes direct constraint on loop length
G4P	Percentage (between 0 and 100%)	Number of windows (100 bp size, 20 bp step) containing at least one G-quadruplex (without loop constraint)	Returns potential rather than putative locations; imposes constraint on total length
QFP	Matched sequence motif	Overall length of G-quadruplex 100	Less stringent than regular expression criterion; imposes constraint on total length