Biochimie 151 (2018) 115–118

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Short communication

G4RNA screener web server: User focused interface for RNA G-quadruplex prediction

Jean-Michel Garant, Jean-Pierre Perreault*, Michelle S. Scott**

RNA Group/Groupe ARN, Département de Biochimie, Faculté de Médecine et des Sciences de la Santé, Pavillon de Recherche Appliquée au Cancer, Université de Sherbrooke, 3201 rue Jean-Mignault, Sherbrooke, Québec, J1E 4K8, Canada

ARTICLE INFO

Article history:
Received 6 March 2018
Accepted 4 June 2018
Available online 6 June 2018

Keywords:
RNA G-quadruplex
Webserver
Bioinformatic predictor

ABSTRACT

Though RNA G-quadruplexes became a focus of study over a decade ago, the main challenge associated with the identification of new potential G-quadruplexes remains a bottleneck step. It slows the study of these non-canonical structures in nucleic acids, and thus the understanding of their significance. The G4RNA screener is an accurate tool for the prediction of RNA G-quadruplexes but its deployment has brought to light an issue with its accessibility to G-quadruplex experts and biologists. G4RNA screener web server is a platform that provides a much needed interface to manage the input, parameters and result display of the main command-line ready tool. It is accessible at http://scottgroup.med.usherbrooke.ca/G4RNA_screener/.

1. Introduction

RNA functions are closely associated to their structural features [1]. While most structures rely on Watson-Crick base pairs and duplex formation, G-quadruplexes (G4) are tetrahelices relying on Hoogsteen base pairing of guanines [2]. Intramolecular RNA G4 are very stable *in vitro* [3], observable in cells [4,5], and have been associated with several functions related to post-transcriptional gene expression regulation [2,6,7].

The reliance of the G4 structures on a local high G content and the resolution of the first G4 structures with the strand of the helix consisting of stretches of G brought the postulation of the motif $G_{\geq 3}N_{1\text{-}7}G_{\geq 3}N_{1\text{-}7}G_{\geq 3}N_{1\text{-}7}G_{\geq 3}$ as the requirement to observe a G4 [8]. This motif was used in the first evaluations of the number of G4 in the human genome yielding successful observations of several potential G4 [9–12]. However, recent years were prolific in the reports of new RNA G4 with an increasing proportion not matching the previously postulated pattern [13–16]. Significant efforts have been deployed since then to adjust the motif in order to include the new patterns of G4 described, but the usage of a more inclusive motif increases the risks of false positive discovery [17]. This issue was addressed by a two independent approaches to discriminate potential G4 using a scoring system [18,19].

These approaches were based on expert knowledge and considered a limited number of observed structures that are now assumed to depict an incomplete picture of all G4 conformation possibilities. Such a strategy is not suitable in a discovery driven approach aiming to allow new conformations to be picked up by the prediction tool. We chose to let the data drive the predictions and implemented a minimal machine learning model to train itself in the recognition of G4 prone sequences based on the sequences found in G4RNA database [20]. The result, G4NN, is an artificial neural network demonstrated to have excellent predictive power and to be especially efficient at discarding randomly selected transcripts [21].

Since other previous approaches also provided satisfactory predictions, valuable information and insight on the sequence, we included in G4RNA screener the scoring systems that were not reliant on pre-defined nucleotidic motifs. G4Hunter (G4H) and the consecutive G over consecutive C (cGcC) scoring systems [18,19] are available along G4NN in G4RNA screener.

G4RNA screener was originally released in its command line form [21]. However, because most users are not familiar with this interface, our latest endeavor has since been to produce a graphical interface which facilitates access to G4RNA screener to a wider audience.

* Corresponding author.
** Corresponding author.
E-mail addresses: jean-pierre.perreault@usherbrooke.ca (J.-P. Perreault), michelle.scott@usherbrooke.ca (M.S. Scott).

https://doi.org/10.1016/j.biochi.2018.06.002
0300-9084/

2. Methods and implementation

The stand-alone command-line program has been thoroughly described [21]. It allows the analysis of large sequences efficiently. G4RNA screener is written in Python and is easily importable. We integrated it in a Django web environment which acts as a bridge between the user's browser and the maintained instance of the program on the server. Therefore, an analysis performed on the webserver will use the same version of G4RNA screener as the stand-alone command line version available to download (http://gitlabscottgroup.med.usherbrooke.ca/J-Michel/g4rna_screener). The time between submission to the server and page refresh to display the results is a matter of seconds or minutes depending on the size of the sequences and the parameters selected by the user.

The usage of a server for computation provides reproducibility and reliability for the user at the cost of limiting the size of submission to 20 000 characters in plain text or 30 kBytes FASTA file. Larger analysis should be conducted on the user resources where the limitations will be dependent on the hardware. One can expect to process ~60 000 nt/min on a 3rd generation Intel i7 desktop processor or ~70 000 nt/min on a 4th generation Intel Laptop processor using the default parameters. The output of the analyses is highly dependent on user input and its format. A parser browses the description lines of the input FASTA file to retrieve the available information. This information is used to provide more detailed results such as the chromosomal position of the analyzed sequence and cross-reference IDs if available in the UCSC database.

Results are sent to the user and displayed in a Javascript datatable with sorting and filtering options. Sequences scored above the provided thresholds have their scores highlighted in the results table for fast identification. The results table can be downloaded either in Excel spreadsheet format for subsequent consultation or in tab delimited text format (.csv) suitable for many tools and facilitating further analysis. Further documentation is accessible online in the web server help page and in the repository manual.

3. Results and discussion

Prediction of G-quadruplexes is an active area of development with multiple tools developed in the last decade, four of them featuring a web interface. Unfortunately, amongst these four webservers, Quadfinder [22] and QuadPredict [23] are currently unavailable and listed as deprecated by OMICtools [24], while the other two, QGRS mapper 2 [9] and QuadBase2's TetraplexFinder [12] are redundant, both searching for motifs. The TetraplexFinder offers a better user experience since QGRS mapper's batch submission does not support multi-FASTA format and requires the user to copy and paste each sequence. QuadBase2's TetraplexFinder can analyze more sequences and at a faster rate than G4RNA screener as expected for a motif matching algorithm. The recent Quadron [25] is an improvement on classical motif searching as it uses stringent motif matching and gradient boosting machines trained on high-throughput detection of G4 to reduce the false positive and false discovery rates of motif matching. Quadron is a very relevant tool that includes a graphic interface but requires a lengthy installation.

Unfortunately, all the previously mentioned tools are focused on the discovery of DNA G4 which is often extrapolated to RNA without thorough comparisons. The available tools focusing on RNA are the cGcC score [18] and RNAfold v2.1.0+ [10]. However, RNAfold does not support G4 on its webserver and also relies on motif matching. As stated previously, G4RNA screener combines the available tools that are not relying on motif search; G4NN [21], cGcC score [18] and G4Hunter [19] none of them having a graphical user interface. G4RNA screener web server provides a time efficient and reliable way to predict G4 folding in RNA sequences. Its intuitive interface manages the analysis parameters and input of G4RNA screener (Fig. 1A).

Default values are recommended for an optimal usage of the G4NN score (Fig. 1B) [21]. Since its training was performed on a set of sequences with lengths distributed around 60 nucleotides (nt), its best performances are obtained at window size 60 nt. However, users are free to increase the size of the analysis window to favor the retrieval of large potential G4 by the tool; inversely, a smaller window will disfavor large G4. The usage of cGcC score is less dependent on the length of the sequence analyzed and offers good performances when dealing with large portions of sequences since it was designed to consider the RNA context in the vicinity of a potential G4 [18]. In order to use the cGcC score in its original usage, the window size can be adapted to cover regions that would represent an appropriate folding space, i.e. the sequence of an internal ribosome entry site, an entire 5′UTR, etc. The G4H score was designed to identify potential G4 in DNA using windows smaller than 60 nt (~30 nt) [19], but it was shown to have good predictive power in RNA using 60 nt windows [21]. Once again, the window size can be adapted to retrieve the behavior of G4H as it was originally reported.

The step size parameter regulates the overlap length between each window by defining the movement length along the sequence between consecutive windows. Therefore, the step size defines the resolution of the analysis (Fig. 1B). It is set to 10 nt by default to reduce the computational burden while providing an adequate resolution. Reducing the step size on sequence regions with ambiguous scores can provide more insights but very low step size means a larger overlap between the windows and, generally, a very low impact on the scores of each consecutive window.

The sample FASTA that is provided as an input example in the web interface (Fig. 1C) was designed so that users can experiment with different description line styles, i.e. the FASTA header conventions supported by G4RNA screener. Most of the display fields available are reliant on the information provided in the description lines (Fig. 1D). Whether it is linked to identification, annotation or chromosomal position, G4RNA screener relies on the user input to collect information. G4RNA screener supports two main structures of description lines; the UCSC refGene description line and the Ensembl transcript description line. Users can take advantage of the patterns to supply the information needed for their subsequent use and the original description line can be retrieved in the description field of the results (Fig. 1E).

All analyzed windows are displayed in the result table (Fig. 1F) which is suitable to compare the scores of a single region to the overall gene. Highly scored windows can be identified by sorting the windows by scores. Multiple sorting is available and useful to analyze multiple sequences individually in a single submission. The table can be downloaded in a spreadsheet for further consultation and interpretation or in tab delimited values in a text file. Any tab delimited value file of chromosomal position with one of the three scores can easily be used to generate a bedgraph file. The visualization of a bedgraph file in a genome browser allows a quick identification of G4 hotspots at a glance (Fig. 2) [26].

4. Conclusion

G4RNA screener web server constitutes our latest endeavor in supporting biologists to identify potential RNA G4. To improve the accessibility of the tool, we provided the user interface in a web page which both eliminates the need to install the tool locally and the management of updates, dependencies as well as reducing the computational burden. We used our own experience as experimentalists to identify the needs of the users, hence the variety of the parameters and options. Nonetheless, users should be aware

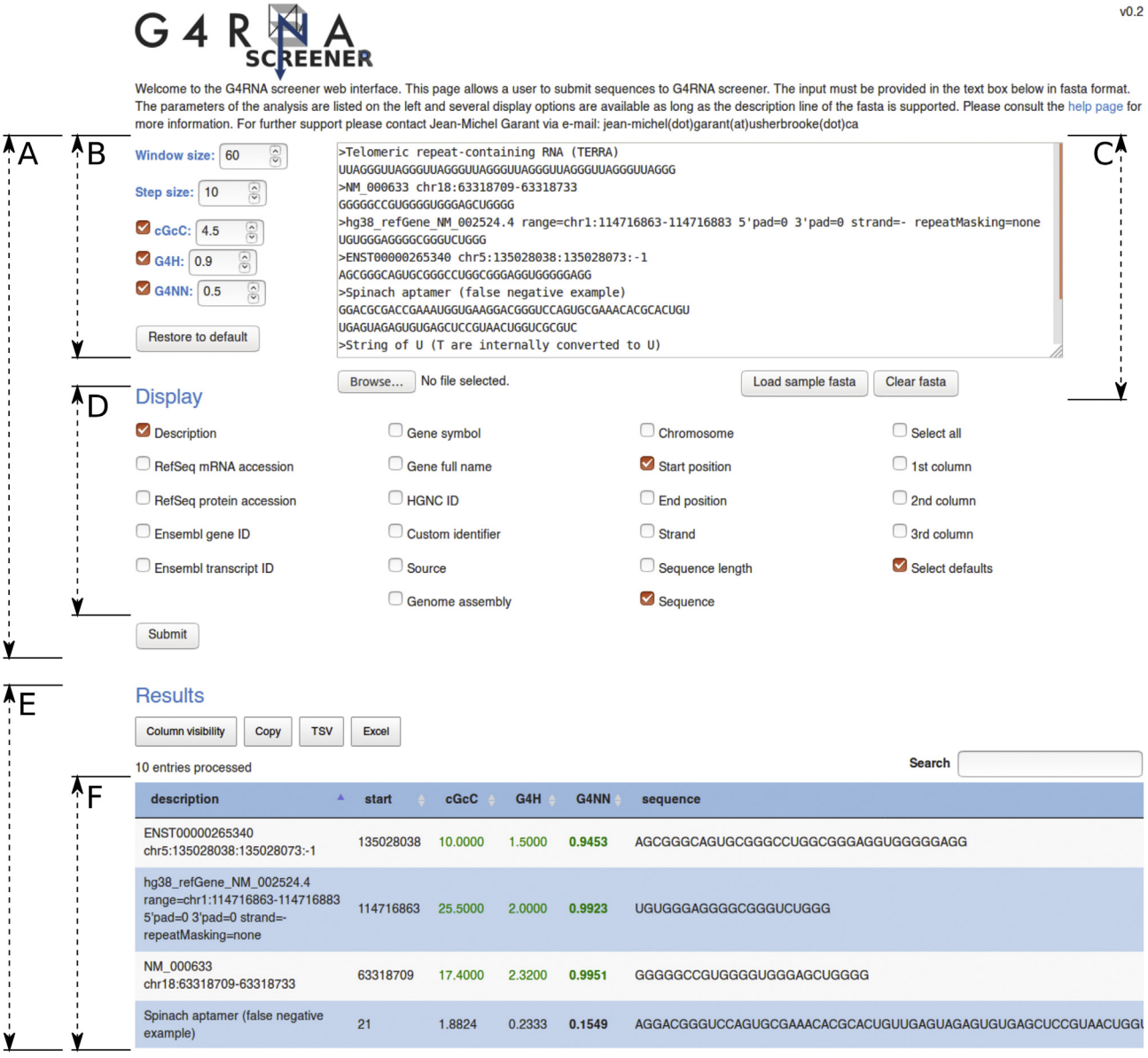

Fig. 1. Screen capture of a typical usage of G4RNA Screener. A) The input form includes 3 sections; B) Analysis parameters, C) FASTA input, D) Display options. E) The results section is displayed following a form submission. F) The analysis output is displayed in a dynamic table.

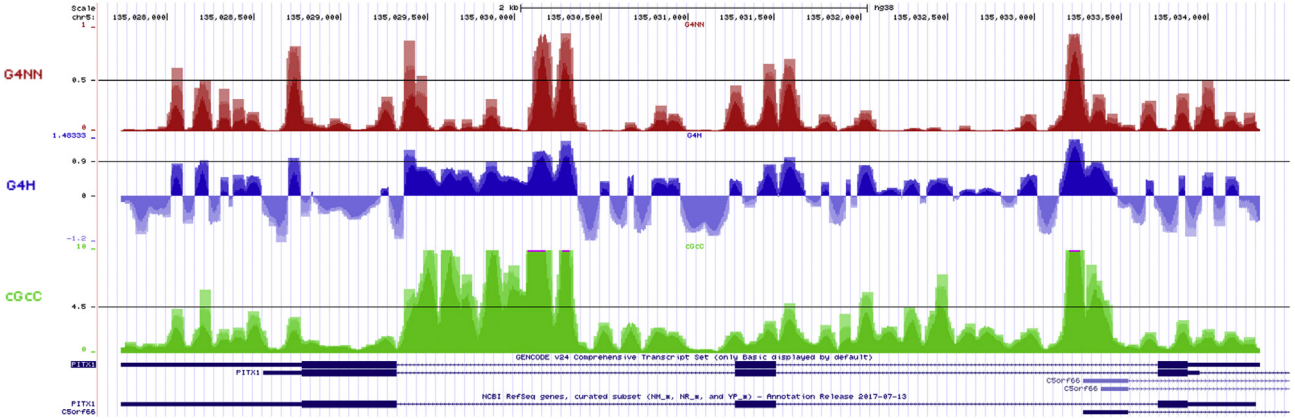

Fig. 2. Bedgraph visualization using the UCSC genome browser of the PITX1 gene scored using G4RNA screener; G4NN in red, G4Hunter in blue, cGcC in green. Each default threshold is shown as a horizontal black line.

that the command line based tool grants more flexibility of analysis.

G4RNA screener is still under active development with potential important impacts pertaining to user experience. The optimization of the computing efficiency is planned with the intent to improve the limits of the web server. G4RNA screener is accessible at http://scottgroup.med.usherbrooke.ca/G4RNA_screener/.

Conflicts of interest

The authors declare no conflict of interest.

Funding

JMG is the recipient of a Natural Sciences and Engineering Research Council of Canada (NSERC) graduate scholarship. MSS holds a Fonds de Recherche du Québec – Santé (FRQS) Research Scholar Junior 2 Career Award. JPP holds the Chaire de recherche de l'Université de Sherbrooke en Structure et Génomique de l'ARN. The project was supported by funding from the Fonds de Recherche du Québec Nature et Technologies (FRQ-NT) (to JPP) and NSERC (RGPIN-2018-05412) (to MSS).

Acknowledgement

The authors are grateful to members of their research groups for helpful comments and critical proof reading.

References

[1] P.A. Sharp, The centrality of RNA, Cell 136 (2009) 577–580, https://doi.org/10.1016/j.cell.2009.02.007.
[2] S. Rouleau, R. Jodoin, J.-M. Garant, J.-P. Perreault, RNA G-quadruplexes as Key Motifs of the Transcriptome, SpringerLink, Springer, Berlin, Heidelberg, 2017, pp. 1–20, https://doi.org/10.1007/10_2017_8.
[3] S. Pandey, P. Agarwala, S. Maiti, Effect of loops and G-Quartets on the stability of RNA G-quadruplexes, J. Phys. Chem. B 117 (2010) 6896–6905, https://doi.org/10.1021/jp401739m.
[4] G. Biffi, M. Di Antonio, D. Tannahill, S. Balasubramanian, Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells, Nat. Chem. 6 (2014) 75–80, https://doi.org/10.1038/nchem.1805.
[5] C.K. Kwok, S. Balasubramanian, Targeted detection of g-quadruplexes in cellular RNAs, Angew. Chem. Int. Ed. 54 (2015) 6751–6754, https://doi.org/10.1002/anie.201500891.
[6] P. Agarwala, S. Pandey, S. Maiti, The tale of RNA G-quadruplex, Org, Biomol. Chem. 13 (2015) 5570–5585, https://doi.org/10.1039/C4OB02681K.
[7] D. Rhodes, J.H. Lipps, G-quadruplexes and their regulatory roles in biology, Nucleic Acids Res. (2015). https://doi.org/10.1093/nar/gkv862.
[8] J.L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human genome, Nucleic Acids Res. 33 (2005) 2908–2916, https://doi.org/10.1093/nar/gki609.
[9] O. Kikin, L. D'Antonio, P.S. Bagga, QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences, Nucleic Acids Res. 34 (2006) W676–W682, https://doi.org/10.1093/nar/gkl253.
[10] R. Lorenz, S.H. Bernhart, J. Qin, C.H. z Siederdissen, A. Tanzer, F. Amman, I.L. Hofacker, P.F. Stadler, 2D meets 4G: g-quadruplexes in rna secondary structure prediction, IEEE ACM Trans. Comput. Biol. Bioinf 10 (2013) 832–844, https://doi.org/10.1109/TCBB.2013.7.
[11] C. Menendez, S. Frees, S.P. Bagga, QGRS-H Predictor: a web server for predicting homologous quadruplex forming G-rich sequence motifs in nucleotide sequences, Nucleic Acids Res. 40 (2012) W96–W103, https://doi.org/10.1093/nar/gks422.
[12] P. Dhapola, S. Chowdhury, QuadBase2: web server for multiplexed guanine quadruplex mining and visualization, Nucleic Acids Res. 44 (2016) W277–W283, https://doi.org/10.1093/nar/gkw425.
[13] H. Huang, B.N. Suslov, N.-S. Li, A.S. Shelke, E.M. Evans, Y. Koldobskaya, A.P. Rice, A.J. Piccirilli, A G-quadruplex-containing RNA activates fluorescence in a GFP-like fluorophore, Nat. Chem. Biol. 10 (2012) 686–691, https://doi.org/10.1038/nchembio.1561.
[14] H. Martadinata, A.T. Phan, Formation of a stacked dimeric G-Quadruplex containing bulges by the 5′-terminal region of human telomerase rna (hTERC), Biochemistry (Mosc.) 53 (2014) 1595–1600, https://doi.org/10.1021/bi4015727.
[15] R. Jodoin, L. Bauer, J.-M. Garant, A.M. Laaref, F. Phaneuf, J.-P. Perreault, The folding of 5′-UTR human G-quadruplexes possessing a long central loop, RNA 20 (2014) 1129–1141, https://doi.org/10.1261/rna.044578.114.
[16] F. Bolduc, J.-M. Garant, F. Allard, J.-P. Perreault, Irregular g-quadruplexes found in the untranslated regions of human mRNAs influence translation, J. Biol. Chem. 291 (2016) 21751–21760, https://doi.org/10.1074/jbc.M116.744839.
[17] C.K. Kwok, G. Marsico, A.B. Sahakyan, V.S. Chambers, S. Balasubramanian, rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome, Nat. Methods 13 (2016) 841–844, https://doi.org/10.1038/nmeth.3965.
[18] J.-D. Beaudoin, R. Jodoin, J.-P. Perreault, New scoring system to identify RNA G-quadruplex folding, Nucleic Acids Res. 42 (2014) 1209–1223, https://doi.org/10.1093/nar/gkt904.
[19] A. Bedrat, L. Lacroix, J.-L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter, Nucleic Acids Res. 44 (2016) 1746–1759, https://doi.org/10.1093/nar/gkw006.
[20] J.-M. Garant, M.J. Luce, M.S. Scott, J.-P. Perreault, G4RNA: an RNA G-quadruplex database, Database 2015 (2015) 1–5, https://doi.org/10.1093/database/bav059.
[21] J.-M. Garant, J.-P. Perreault, M.S. Scott, Motif independent identification of potential RNA G-quadruplexes by G4RNA screener, Bioinformatics 33 (2017) 3532–3537, https://doi.org/10.1093/bioinformatics/btx498.
[22] V. Scaria, M. Hariharan, A. Arora, S. Maiti, Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences, Nucleic Acids Res. 34 (2006) W683–W685, https://doi.org/10.1093/nar/gkl299.
[23] M.H. Wong, O. Stegle, S. Rodgers, L.J. Huppert, A toolbox for predicting g-quadruplex formation and stability, J. Nucleic Acids 2010 (2010), https://doi.org/10.4061/2010/564946.
[24] V.J. Henry, A.E. Bandrowski, A.-S. Pepin, B.J. Gonzalez, A. Desfeux, OMICtools: an informative directory for multi-omic data analysis, Database J. Biol. Databases Curation 2014 (2014), https://doi.org/10.1093/database/bau069.
[25] A.B. Sahakyan, G. Marsico, M.D. Antonio, S. Balasubramanian, T. Santner, V.S. Chambers, Machine learning model for sequence-driven DNA G-quadruplex formation, Sci. Rep. 7 (2017) 14535, https://doi.org/10.1038/s41598-017-14017-4.
[26] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006, https://doi.org/10.1101/gr.229102.