OXFORD

Sequence analysis

G4Hunter web application: a web server for G-quadruplex prediction

Václav Brázda[1,*], Jan Kolomazník[2], Jiří Lýsek[2], Martin Bartas[1,3], Miroslav Fojta[1], Jiří Šťastný[2,4] and Jean-Louis Mergny[1,5]

[1]Department of biophysical chemistry and molecular oncology, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno 612 65, Czech Republic, [2]Department of Informatics, Mendel University in Brno, Brno 613 00, Czech Republic, [3]Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava 710 00, Czech Republic, [4]Brno University of Technology, Technicka 2896/2, 61969 Brno, Czech Republic and [5]ARNA Laboratory, Inserm U1212, CNRS UMR5320, IECB, Université de Bordeaux, Pessac, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 25, 2018; revised on January 8, 2019; editorial decision on January 31, 2019; accepted on February 4, 2019

Abstract

Motivation: Expanding research highlights the importance of guanine quadruplex structures. Therefore, easy-accessible tools for quadruplex analyses in DNA and RNA molecules are important for the scientific community.

Results: We developed a web version of the G4Hunter application. This new web-based server is a platform-independent and user-friendly application for quadruplex analyses. It allows retrieval of gene/nucleotide sequence entries from NCBI databases and provides complete characterization of localization and quadruplex propensity of quadruplex-forming sequences. The G4Hunter web application includes an interactive graphical data representation with many useful options including visualization, sorting, data storage and export.

Availability and implementation: G4Hunter web application can be accessed at: http://bioinformatics.ibp.cz.

Contact: vaclav@ibp.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Important roles are emerging for local DNA structures in the regulation of basic biological processes (Brázda *et al.*, 2011; Todolli *et al.*, 2017). Sequences that can form G-quadruplex (G4) structures are widespread in a variety of genomes (Huppert and Balasubramanian, 2005; Murat and Balasubramanian, 2014; Todd *et al.*, 2005). G4 structures often exhibit high structural stability under physiological conditions and are important in telomere biology (De Cian *et al.*, 2008; Zimmermann *et al.*, 2014), protein recognition (Brázda and Coufal, 2017; Brázda *et al.*, 2018, 2014; Vasilyev *et al.*, 2015), transcription (Siddiqui-Jain *et al.*, 2002), translation and RNA maturation (Wieland and Hartig, 2007), replication and genomic stability (Cheung *et al.*, 2002; Paeschke *et al.*, 2011) and replication origin definition (Comoglio *et al.*, 2015; Valton *et al.*, 2014). G4 sequences can be predicted *in silico* and several tools have been developed (Garant *et al.*, 2018; Hon *et al.*, 2017; Huppert and Balasubramanian, 2005; Kikin *et al.*, 2006; Wong *et al.*, 2010). A newly developed and validated algorithm, G4Hunter, overcomes several limitations of previous algorithms (Bedrat *et al.*, 2016).

G4Hunter is a powerful and widely used tool for G4 prediction which takes into account G-richness and G-skewness of a DNA or RNA sequence and provides a quadruplex propensity score. However, three major limitations of the G4Hunter have become evident. (i) The current implementation requires advanced computational expertise and software packages including Python v. 2.7 and libraries NumPy, Matplotlib and Biopython; (ii) The software runs in the command line only; (iii) results are exported in two files in text format without any additional tools for direct postprocessing or visualization.

Here, we present 'G4Hunter web application', a new and optimized algorithm, expanding on the original G4Hunter Python

code embedded into a platform-independent and easy-to-use web-based graphical interface, freely available at http://bioinformatics.ibp.cz.

2 Features

Our implementation is based on the algorithm published on GitHub https://github.com/AnimaTardeb/G4-Hunter (Bedrat *et al.*, 2016). We modified it to our server-based platform developed originally for Palindrome analysis (Brázda *et al.*, 2016). For this application, we completely rewrote the back-end and the front-end of our web-server to broaden and optimize computational and visualization performance.

2.1 Workflow and implementation

The original G4Hunter requires the user to install Python and Python's libraries Biopython, Matplotlib and NumPy, runs only in console and stores results in a text file. The basic input parameters and result of our new application are the same; the user specifies input sequence(s), window size and G4Hunter score threshold, and outputs the sequences and scores identified. Compared to the original Python program, our web-based application has a user-friendly interface and rich graphical presentation of results. It also shows a heatmap of quadruplex forming sequences and provides statistical information. Another advantage is the ability to analyze multiple sequences at once. Results can also be exported as a text file for further processing. The G4Hunter web application runs on a central web server and clients/visitors can execute analyses using server resources, sequences and results are stored in a relational database. A compatible web browser with JavaScript support is required. Although user requests are queued and processed one at a time, the application benefits from multi-core processor architectures because the analysis itself is performed by Java execution framework and the algorithm is separated into four independent steps (determine the score for each nuclide, precalculate sum, calculate scores and aggregation) that are computed in parallel. Our application is implemented as a Single-Page-Application with REST API back-end. This configuration allows us to use the back-end as a stand-alone computational server employed by custom scripts for specific analyses. Front-end client application is made with Vue.js (https://vuejs.org/) framework. Back-end is a Java application based on Spring framework (https://spring.io/). The server currently runs on 4 threads.

2.1.1 G4Hunter web application input and analysis

Sequences can be imported using NCBI ID, with the ability to download multiple sequences from NCBI multiple import. Another option is to upload a local FASTA or txt file and/or directly paste sequences from the clipboard. The user can sort sequences according to all parameters and can add tags to organize sequence sets. Standard IUPAC nucleotide codes are supported. Sample sequences are available to test the server. The user selects one or more sequences and sets the parameters for analysis. Recommended parameters for Window size (25) and score Threshold (1.4) are set as default (window size 10-100 nucleotides and threshold 0.8-4.0 are available, the latter will only retrieve pure G-runs).

2.1.2 G4Hunter web application output

Results are displayed using AJAX technology. If more than one sequence is analyzed, results are shown as individual tabs. The user interface is designed for intuitive visualization and browsing of results. Below the sequence name is a heatmap that divides the sequence and displays the number of G4-forming sequences in each segment. The number of results is marked by the intensity of red color. The heatmap can be used to filter results in selected segments. Below this is basic information including settings, number and frequency of results, export options and sequence information, including length and GC content. The sequence browser component displays the nucleotide sequences of the results and shows a cut-out of bases which fits the screen/browser window. The sequences corresponding to the analysis parameters are marked by colors (G in red—the longer the G-track, the brighter the intensity, C in blue). Position in the sequence, length, sequence, score chart and G4Hunter score are shown. The aggregated results are shown primarily; separated results are displayed using the magnifier icon. A typical analysis output is shown in Supplementary Material S1.

2.2 Output formats

G4Hunter web application gives three result formats. First, the graphical representation described above. Second, concatenated sequences and third, unaggregated sequences in CSV files. The structure of CSV files is shown in Figure 1 and contains POSITION in sequence, LENGTH of the longest continuous sequence with G4Hunter scores above threshold, its SCORE and the part of the sequence. The SUB_SCORE shows scores for each window position inside the concatenated sequence. These CSV files can be downloaded from the main results window and/or from the stored results tabs for follow-up analyses.

```
"ID" "POSITION" "LENGTH" "SCORE" "SEQUENCE" "SUB_SCORE"
"1"  "926"      "27"     "1.30"  "AAG…GTT"  "1.40, 1.40, 1.40"
"2"  "935"      "25"     "1.40"  "ACG…CGG"  "1.40"
"3"  "937"      "25"     "1.40"  "GGG…GCT"  "1.40"
"4"  "6819"     "40"     "-0.82" "CCA…CCT"  "-1.44, …, -1.48"
```

Fig. 1. CSV export

3 Validation

To compare the performance and accuracy of G4Hunter web application, we analyzed several identical sequences in each version. Both implementations return the same results. However, thanks to the new architecture of our application and parallel processing, the web version is more than 10-times faster and allows analyses of multiple sequences with a system-independent modern graphical environment.

4 Conclusions

We developed a web version of the G4Hunter application with a user-friendly GUI and improved output options including graphic representation. Our web-server allows detailed analyses of nucleic acid sequences and adds basic information and broad visualization options with sorting tools that allow quick and effective searching for target information from G4Hunter analyses. This web version of the G4 algorithm allows rapid and effective analyses of various

nucleic acid sequences and will be useful for researchers in the field.

Acknowledgements

Funding

References

Bedrat,A. *et al*. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res*., **44**, 1746–1759.

Brázda,V. *et al*. (2011) Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol*., **12**, 33.

Brázda,V. *et al*. (2014) DNA and RNA quadruplex-binding proteins. *Int. J. Mol. Sci*., **15**, 17493–17517.

Brázda,V. *et al*. (2016) Palindrome analyser—a new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun*., **478**, 1739–1745.

Brázda,V. *et al*. (2018) The amino acid composition of quadruplex binding proteins reveals a shared motif and predicts new potential quadruplex interactors. *Molecules*, **23**, 2341.

Brázda,V. and Coufal,J. (2017) Recognition of local DNA structures by p53 protein. *Int. J. Mol. Sci*., **18**, 375.

Cheung,I. *et al*. (2002) Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nat. Genet*., **31**, 405–409.

Comoglio,F. *et al*. (2015) High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep*., **11**, 821–834.

De Cian,A. *et al*. (2008) Targeting telomeres and telomerase. *Biochimie*, **90**, 131–155.

Garant,J.-M. *et al*. (2018) G4RNA screener web server: user focused interface for RNA G-quadruplex prediction. *Biochimie*, **151**, 115–118.

Hon,J. *et al*. (2017) pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*, **33**, 3373–3379.

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

Kikin,O. *et al*. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res*., **34**, W676–W682.

Murat,P. and Balasubramanian,S. (2014) Existence and consequences of G-quadruplex structures in DNA. *Curr. Opin. Genet. Dev*., **25**, 22–29.

Paeschke,K. *et al*. (2011) DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell*, **145**, 678–691.

Siddiqui-Jain,A. *et al*. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl Acad. Sci. USA*, **99**, 11593–11598.

Todd,A.K. *et al*. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res*., **33**, 2901–2907.

Todolli,S. *et al*. (2017) Contributions of sequence to the higher-order structures of DNA. *Biophys. J*., **112**, 416–426.

Valton,A.-L. *et al*. (2014) G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J*., **33**, 732–746.

Vasilyev,N. *et al*. (2015) Crystal structure reveals specific recognition of a G-quadruplex RNA by a β-turn in the RGG motif of FMRP. *Proc. Natl Acad. Sci. USA*, **112**, E5391–E5400.

Wieland,M. and Hartig,J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem. Biol*., **14**, 757–763.

Wong,H.M. *et al*. (2010) A toolbox for predicting G-quadruplex formation and stability. *J. Nucleic Acids*, **2010**, 1.

Zimmermann,M. *et al*. (2014) TRF1 negotiates TTAGGG repeat-associated replication problems by recruiting the BLM helicase and the TPP1/POT1 repressor of ATR signaling. *Genes Dev*., **28**, 2477–2491.