Deliverable 3.3 — Gene Regions and G-quadruplex Structures: Using GenData2020 data model with a G4 Predictor

Sergio Greco, Giuseppe Tradigo, Pierangelo Veltri

DIMES - University of Calabria

1 Introduction

DNA is a long polymer being famous for its double-helix form at the lower level, and for its chromosome packaging at higher levels of detail. Nonetheless DNA exists in many possible conformations, including A-DNA, B-DNA and Z-DNA forms. B-DNA is the most common form found in cells. Non-B DNAs comprise of tetraplex (G-quadruplex), left-handed Z-DNA, and others. Several recent publications have provided significant evidence that non-B DNA structures may play a role in DNA instability and mutagenesis, leading to both DNA rearrangements and increased mutational rates, which are a hallmark of cancer related diseases. Studying the structure conformation and probability of non-B DNA structure, may help in studying diseases as well as designing of new drugs. Nevertheless, even if there are some examples of prediction tools, the topic of designing efficient prediction algorithms and tools for G-quadruplex prediction is still in its infancy.

As a contribution in this new area, we present preliminary results and statistics obtained by using the state of the art software tools able to predict G-quadruplex DNA conformations starting from the primary sequence. We used existing tools as well as known structures to define the state of the art and the current value of prediction tools.

Results of this works have been used to define a new prediction tool that is available on line. We also report about the possible integration of the tool with the GenData2020 system. The contents of this report have been published in [1].

G-quadruplexes consist of four-stranded nucleic acids and are described as presenting a high-order secondary DNA structure. In recent years G-quadruplexes structures have drawn attention regarding their potential use in anti-cancer clinical therapies. These guanine-rich (G-rich) sequences exist ubiquitously in significant regions of the eukaryotic genome, such as in the promoter regions of several oncogenes and in the telomeres, the terminal portions of the chromosomes. It has been reported that telomerase is active and up-regulated in approximately 85% of tumor cells, which leads to telomere elongations and contribution to cancer cell immortalization. Thus, the telomeric G-quadruplex has been considered to be a potentially effective anti-tumor target. The formation of

a G-quadruplex would result in the inhibition of telomerase activity and would thereby terminate telomere maintenance. Research aimed at the stabilization of the G-quadruplex structure of certain sequences and efficient inhibition of telomerase activity represents a rising field of research in anti-cancer drug design and development [2]. Many of the reported G-quadruplex ligands evaluated in biological essays have shown in vitro activities, including telomerase inhibition, oncogene down-regulation and suppression of cancer cell proliferation [3].

The target of this report is to assess the impact and quality of software tools that can recognize a sequence forming a G-quadruplex. This may help to establish which molecules (ligands) are able to stabilize that structure. These software tools are known as prediction tools or simply predictors. DNA structure prediction is usually implemented by determining its secondary structure first. Secondary structure prediction depends primarily on the coupling interactions between bases and bases stacking. In fact many molecules can have different three-dimensional structures. Therefore understanding all aspects of G-quadruplex structures is essential in order to create better DNA folding models and algorithms. These structures are characterized by the superposition of planar quartets (G-tetrad) composed of four guanines that interact with Hoogsteen-type hydrogen bonds. Hoogsteen base pairing refers to the formation of of hydrogen bonds between nucleo-bases from different portions of the DNA strand. The G-tetrad, or guanine core, is formed by superposition of three tetrads and supported by four filaments of the phosphodiester skeleton. The loops are portions of the nucleotide sequence that connect the four filaments of the core. The G-quadruplexes structures are further stabilized by ionic interactions established between the oxygen in position 6 of the bases guanine and monovalent metal cations such as sodium (Na^+) and especially potassium (K^+) settled in the center of the structure.

Prediction tools are designed by considering the conformation and chemical characteristics of the DNA polymer. We here report some of the main key features to recognize a G-quadruplex structure and all its possible conformations, according to literature: (*i*) hydrogen bond type Hoogsteen [8]; (*ii*) hydrogen bonds can be clockwise or anticlockwise [10]; (*iii*) type of conformation of the base guanine, *syn* or *anti* [12]; (*iv*) metallic monovalent cations (most commons being Na^+ or K^+) [12]; (*v*) orientation filaments (all parallel, 3 parallel and 1 not, parallel adjacent, all antiparallel) [8, 12]; (*vi*) loop length, sequence and type [8, 10]. These conformation features are in a chemical dynamic equilibrium so that they compete in giving a 3d structure to the G-quadruplex polymer. Being able to design software tools able in identifying such features may give the ability to predict the structure of a G-quadruplex conformation. In [4, 5] there is a detailed description of the G-quadruplex structure.

2 Related Works

Non-B DNA structures have being attracting the interest of researchers and biologists for the role they may play in designing anticancer drugs. Nevertheless, a small number of G-quadruplex structure prediction tools is available, especially considering the number of protein structure prediction tools (see for instance [14]). However there exist some databases of *p*otential G-quadruplex structures and tools able to recognize them. In [6] the 2.0 database version

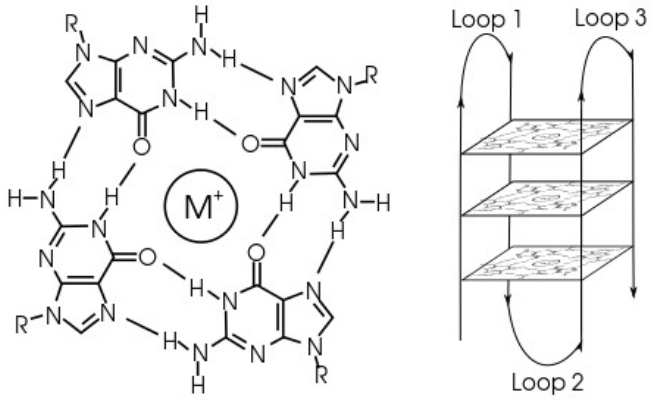

Figure 1: G-quadruplex structure. On the left a tetrad guanine plane is depicted where M^+ is a monovalent cation (typically Na^+ or K^+) and lines among nucleic bases show Hoogteen hydrogen bonds; on the right a parallel adjacent DNA filament folding is depicted, forming three stacked tetrad planes. (*Image source: Wikipedia*)

of predicted non-B DNA-forming motifs is described together with its associated tools. This database provided the first complete annotation of non-B DNA-forming sequence motifs covering the genomes of mammalian organisms including human, mouse, chimpanzee, dog and macaque. In [7], the database of G4 QuadBase genome is described. The Greglist database, described in [8] is a listing of potential G-quadruplex regulated genes containing promoter GQMs from genomes of various species, including humans, mice, rats and chickens. On the web there are some software that can predict whether a given sequence could form a G-quadruplex structure. They are based on different types of parameters and statistical and mathematical models, based on the recognition of the generic pattern in the form of (1) in the input sequence, where $x >= 3$ and N corresponds to any of the letters of a nucleic acid.

$$G_x N_{1-7} G_x N_{1-7} G_x N_{1-7} G_x \tag{1}$$

For instance QGRS Mapper, described in [9], predicts the presence of quadruplex forming G-rich sequences (QGRS) in nucleotide sequences. These presumed G-quadruplex are identified using (1). It consists of four equal length sets of guanines, separated by arbitrary nucleotide sequences, with some limitations such as the sequence length, the length of the loop, the number of tetrads. The scoring system developed by Bagga et al. is very interesting. It evaluates a QGRS for its likelihood to form a stable G-quadruplex. This scoring system is described in greater detail in [10]. It measures a QGRS for its probabilities of forming a stable G-quadruplex and the highest score is given to sequences that are the best candidates to form a stable G-quadruplex. The scoring method uses various principles including: shorter loops are more common than longer loops, G-quadruplexes tend to have loops of almost equal size, and an increased number of tetrads of guanine corresponds to more stable G-quadruplex. One of the criteria used for assigning scores to the QGRS is the uniformity of the

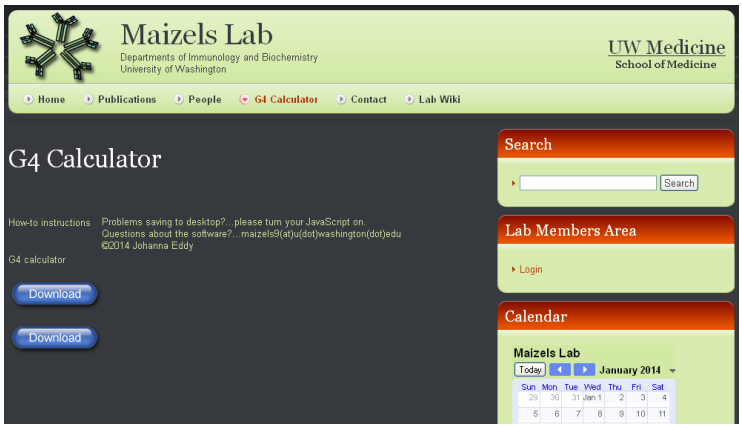

Figure 2: G4P Calculator website.

lengths between the G-runs. The following formula (2) defines the average loops lengths difference (Gavg).

$$G_{avg} = \frac{|L1 - L2| + |L1 - L3| + |L2 - L3|}{3} \tag{2}$$

The G_{score} formula is given by

$$G_{score} = \frac{G_{avg}}{G_{max}} * T(x) \tag{3}$$

in which G_{max} is a normalization constant relative to the length of the G-rich sequence, and $T(x)$ is a function that increases the G_{score} when the number of guanines in each group is greater than two.

3 Prediction Tools

In our work we tested the following predictor tools: *G4P calculator*[1] (see [13]), *QGRS Mapper*[2] (see [9]) and *nBMST*[3] (see [6]). All of them can be invoked via a Web server interface or by downloading a software executable. The websites of the above mentioned tools are reported Figure 2, in Figure 3 and in Figure 4.

G4P Calculator, described in [11], computes G-quadruplex DNA potentials based on guanine runs density (*G-runs*) in a sequence. It evaluates G-runs in a sliding window and calculates the percentage of the searched windows that meets the specified criteria. The G-quadruplex potential is scored as a percentage, thus making it independent from the sequence length. The G4 Calculator software produces results as a text file in tab delimited format containing the following information: (*i*) number of G-runs, (*ii*) number of C-runs (cytosine runs), (*iii*) total number of windows searched, (*iv*) percentage of windows containing G-runs, (*v*) percentage of windows containing C-runs, (*vi*) sum of the

[1] http://depts.washington.edu/maizels9/G4calc.php
[2] http://bioinformatics.ramapo.edu/QGRS/index.php
[3] http://nonb.abcc.ncifcrf.gov/apps/nBMST/default/

Figure 3: QGRS Mapper website.

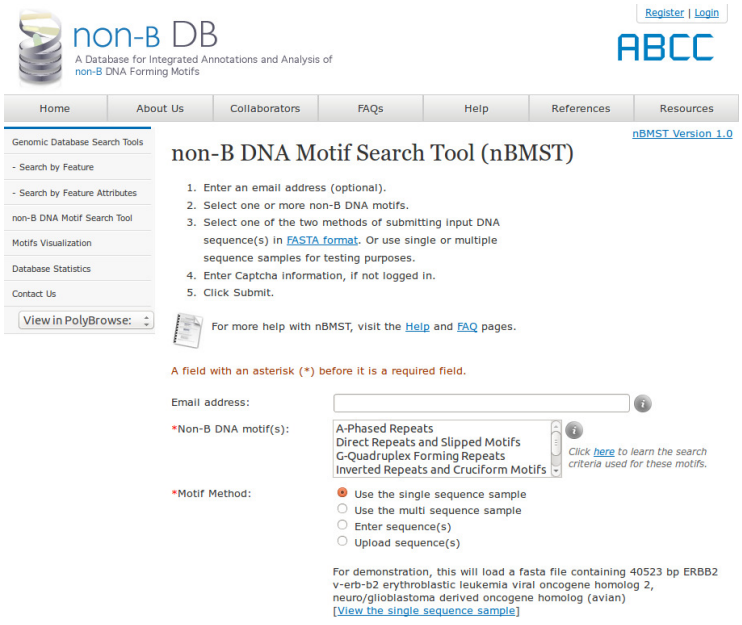

Figure 4: nBMST website.

two percentages, (*vii*) indicator of block of unknown sequence. In this case Ns are removed from the sequence before computing the G4 DNA potential.

QGRS Mapper gives in output the results well-structured and synthesized with the ability to export an excel file. In the results page all the sequences that potentially form a G-quadruplex are shown.

The nBMST tool identifies in the DNA sequences the motif non-B as the G-quadruplexes. The nBMST G-quadruplex recognition algorithm identifies four or more individual G-runs of at least three nucleotides in length. The algorithm requires at least one nucleotide between each run and considers up to seven

nucleotides as spacer, including guanines [6]. nBMST identifies in the DNA sequences the non-B motifs as the G-quadruplexes, the Z-DNA and others. On the results page, you can check the total number of non-B motifs found in the sequence, it display the data in images or in tables and it is also possible to download all the files associated with that job. In this site you can make a registration to have a diary of the work done and the files produced will be retained for about ten months. In the table of results it is possible to find, among others, the following information:

- *Composition* - shows the number of A, C, G, and T nucleotides in the motif sequence. (e.g. $6A/1C/18G/1T$);
- *nIslands* - the number of uninterrupted stretches of guanines that make up the G-quadruplex motif;
- *nRun* - the number of possible non-adjacent runs of 3 or more guanines making up the G-quadruplex motif;
- *maxGQ* - the maximum guanine run length for which 4 or more consecutive runs can be formed in the motif.

All of the three software are capable of recognizing G-quadruplex sequences. Nevertheless none of them is able to predict the 3d structure. We tested the up defined tools on the datasets described in the following section.

4 Datasets

For our experimental phase we set up two datasets containing G-quadruplex structures and non G-quadruplex ones. These datasets allowed us to evaluate the prediction performances of the selected prediction tools. The dataset containing G-quadruplex structures has been created by collecting 160 sequences, 150 contained in the PDB and 10 from the EBI databases, and reducing the structures to avoid redundance to a final number of 82. In fact, while collecting the structures, we noticed that sequences with different names were related to the same sequence and the same structure. In other cases the same sequence lead to different structures; in these cases we kept just one representative sequence for the group.

The controls dataset has been built by randomly selecting 82 sequences from the PDB among the ones not forming G-quadruplex structures. The sequence length distribution has been choosen to be as close as possible to the G-quadruplex ones. The purpose of this dataset is to test the ability of prediction tools with non G-quadruplex structures (i.e. avoiding false positives).

5 Results and Discussion

We invoked the three prediction tools on each of the 82 targets in both the G-quadruplex and non G-quadruplex datasets. Table 1 reports an extract of the top 25 targets in the G-quadruplex dataset, ranked by sequence length (*Seq. Ln*) and by the number of predictors (*n. Pred.*) correctly predicting it as a

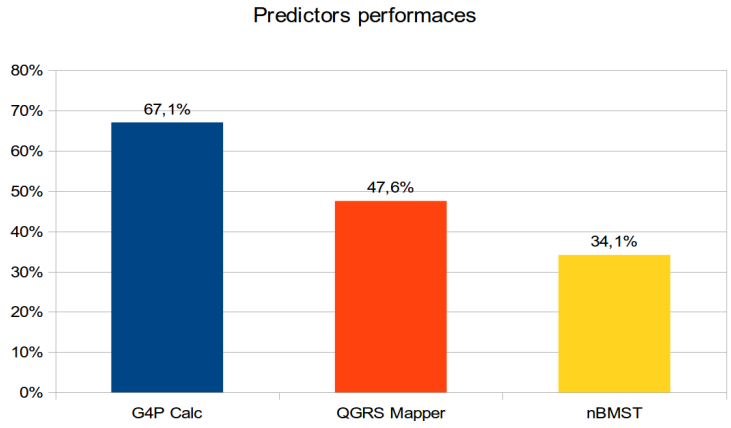

Figure 5: Prediction tools performance over the whole target dataset. The best performing predictor was G4P with more than 67.1% of correctly predicted targets.

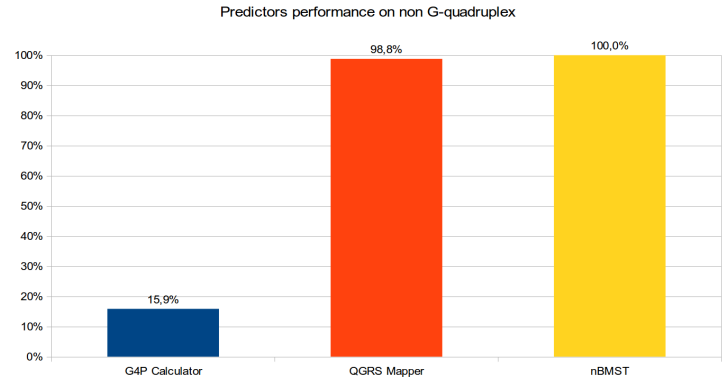

Figure 6: Prediction tools performance over controls dataset. Despite the G4P tool being the top performer in predicting G-quadruplex structures, it is not very precise in identifying non-G-quadruplex ones.

G-quadruplex sequence. The first two columns show the sequence name in the PDB database (*Name*) and the sequence itself (*Sequence*).

Overall, the G4P Calculator was the best in predicting G-quadruplex structures (see Figure 5), with an average score of 67.1% of correctly identified G-quadruplex structures. QGRS Mapper guessed less than a half (47.6%) of the target in the correct class. nBMST had a much worst performance predicting G-quadruplex structures with a score of 34.1%.

Some of the main problems recognizing G-quadruplex targets, for example some PDB sequences with G-run of 2, were related to incompatibility with predictors parameters used during the training phase. Of course these parameters are beyond the authors' control and are discussed here to provide a possible interpretation of the results. For instance, sequences having a length of less than 10 nucleotides and including multiple filaments, are more difficult to identify or

Name	Sequence	n. Pred.	Seq. Ln
230D	GGGGTUTUGGGGTTTTGGGGUUTTGGG	3	27
2JPZ	TTAGGGTTAGGGTTAGGGTTAGGGTT	3	27
2HY9	AAAGGGTTAGGGTTAGGGTTAGGGAA	3	26
2LPW	AAGGGTGGGTGTAAGTGTGGGTGGGT	2	26
2JSL	TAGGGTTAGGGTTAGGGTTAGGGTT	3	25
2A5P	TGAGGGTGGIGAGGGTGGGGAAGG	2	25
2GKU	TTGGGTTAGGGTTAGGGTTAGGGA	3	24
1OZ8	GGAGGAGGAGGAGGAGGAGGAGGA	2	24
2F8U	GGGCGCGGGAGGAATTGGGCGGG	3	23
2JSM	TAGGGTTAGGGTTAGGGTTAGGG	3	23
2KKA	AGGGTTAGGGTTAIGGTTAGGGT	2	23
1XAV	TGAGGGTGGGTAGGGTGGGTAA	3	22
1KF1	AGGGTTAGGGTTAGGGTTAGGG	3	22
2KF7	GGGTTAGGGTTAGGGTTAGGGT	3	22
2KM3	AGGGCTAGGGCTAGGGCTAGGG	3	22
2LOD	GGGATGGGACACAGGGGACGGG	3	22
2LXQ	TAGGGTGGGTTGGGTGGGGAAT	3	22
2O3M	AGGGAGGGCGCTGGGAGGAGGG	3	22
3QXR	AGGGAGGGCGCUGGGAGGAGGG	3	22
2KQG	CGGGCGGGCACGAGGGAGGGT	3	21
2KQH	CGGGCGGGCGCGAGGGAGGGT	3	21
2KYP	CGGGCGGGCGCTAGGGAGGGT	3	21
4DA3	GGGTTAGGGTTAGGGTTAGGG	3	21
1I34	GGTTTTGGCAGGGTTTTGGT	2	21
2L88	GGGGCGGGGCGGGGCGGGGT	3	20
2KZD	AGGGIAGGGGCTGGGAGGGC	3	20
2KZE	AIGGGAGGGICTGGGAGGGC	3	20
2LED	TAGGGCGGGAGGGAGGGAA	3	20
2KOW	TAGGGTAGGGTAGGGTAIGG	2	20
2KPR	GGGTGGGGAAGGGGTGGGT	3	19

Table 1: Table reporting the top 30 length targets from the 82 contained in the target dataset. For each target we report its name, the primary sequence, the number of predictors correctly guessing its G-quadruplex conformation (*n. Pred.*) and the length of the sequence (*Seq. Ln*).

predict as a possible G-quadruplex structure.

The overall performance on the controls dataset (see Figure 6) was particularly odd for the G4P Calculator tool. It scored just 15.9% of correctly predicted non G-quadruplex structure, but it had a very high $80,5\%$ of failures during execution and 3 out of 82 wrongly classified G-quadruplex structures (false positives). Both QGRS Mapper and nBMST performed very well on non G-quadruplex structures with only 1 out of 82 mis-classification for QGRS Mapper (false positive). This is probably due to the too tight parameters with which the tools have been trained.

G4P Calculator was able to recognize the sequences even if they were short and if the G-runs were composed solely from 2 Guanine. QGRS Mapper recognized only G-quadruplex structures with more than 15 nucleotides. nBMST also correctly classified only G-quadruplex structures with more than 16 nucleotides.

For the non G-quadruplex structure prediction the sequence length is unrelated to the prediction performances. Even the shortest ones, got a score of 2 or higher correct predictions.

There is an intrinsic difficulty for these methods to predict properties for short sequences because the algorithms adopted by the tools are contextual and perform better if they can "see" a longer sequence.

6 GQuadruplex prediction tool

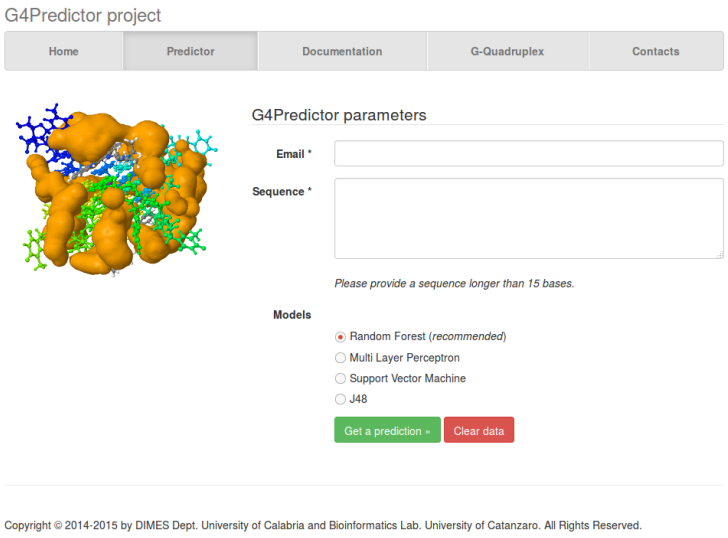

Figure 7: G-quadruplex prediction web-based interface. The input form allows final users to specify a RNA input sequence and to choose a machine learning model to be used to predict potential G4 regions in it.

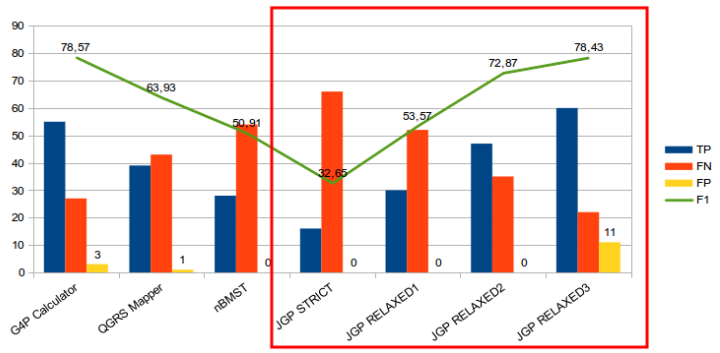

Figure 8: G-quadruplex pattern-based prediction tools performances. The red square shows results for the four predictors developed by us with respect to the prediction tools available in literature. The *JGP RELAXED3* tools shows an F1 index comparable to the best performing tool in literature.

In the context of the GenData2020 project we developed a set of tools for G-Quadruplex (G4) structures prediction. A prediction tool takes an RNA input sequence in the form of a string, where each character represents a nucleic acid, and returns a string of the same length, where each base position is marked as being in a potential G4 sub-structure or not. We worked on a set of pattern-based predictors and another set of machine learning-based tools.

The pattern-based prediction tools, which search for the pattern described

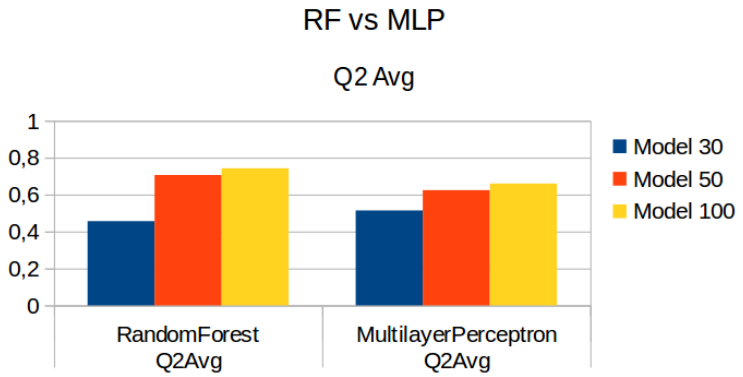

Figure 9: G-quadruplex machine learning-based prediction tools performances. Results are about the two best performing tools predicting G4 sequences locations on three datasets having sequences with lengths up to 30, 50 and 100 bases. The shown measure, EVA Q2, is a count of correctly guessed nucleic acid bases state (G4 or non G4).

in equation (1), with various relaxations (i.e. number of G_xN_{1-7} segments to search). Prediction performances for these tools are depicted in Figure 8, where on the left tools from the literature are reported and on the right (red box) our tools are depicted. The F1 measure shows that the performances of one of our tool (*JGP RELAXED3*) is similar to the best prediction tool described in literature.

The machine learning-based prediction tools have been developed by training a set of machine learning models (e.g. neural network, support vector machine, decision tree) on a dataset of both positive (G4) and negative (normal RNA) examples shown to the model by considering moving windows of adjacent nucleic acids of variable sizes (e.g. 5, 7, 9). The trained prediction models have been packed in a system available on line at [15] which is currently being testing with available datasets. Figure 7 depicts the web based interface form available for users to predict the G4 region locations on a given RNA input sequence, while Figure 9 reports the performance of the two best models (having an optimal window of 15 nucleic acids) tested on three datasets: dataset with sequences up to 30, 50 and 100 nucleic acid bases. The accuracy shown is the EVA Q2 measure, which represents a count of correctly predicted substructure in the base sequence.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<gmqlSchemaCollection name="GLOBAL_SCHEMAS"
  xmlns="http://www.bioinformatics.deib.polimi.it/GMQL/">
  <gmqlSchema name="G4_REGIONS" type="tab">
    <field type="FLOAT">G4</field>
  </gmqlSchema>
</gmqlSchemaCollection>
```

Listing 1: General Tab Format schema file to describe the predicted G-Quadruplex regions to the GMQL package.

```
$ repositoryManagerV1 CreateDS G4_ANN g4.xml g4data.tab
```

Listing 2: Command line to import the G-Quadruplex regions dataset into the GMQL package.

```
chr1 3210 3222 0.81
chr1 3227 3235 0.75
...
```

Listing 3: GMQL Tab data file example. G4 regions are marked with the chromosome, the starting and ending nucleic acid position and a prediction accuracy score.

```
cell                H1-hESC
cell_description    cervical carcinoma
cell_karyotype      cancer
cell_sex            F
cell_tissue         cervix
cell_type           Cell Line
...
```

Listing 4: GMQL Tab metadata file example. It contains description label-value couples for the related dataset.

The proposed system can be used with the GenData2020 data model by importing the predicted G4 regions into the GMQL package. The GenoMetric Query Language (GMQL) operates upon aligned genomic data in a variety of data formats, providing parallel computation in the cloud thus supporting queries over thousands of samples (e.g. ENCODE and TCGA consortia datasets).

The predicted G4 regions, using the simple schema in Listing 1 and imported with the command reported in Listing 2, are represented in the GMQL Tab data format. An extract of the GMQL data file is reported in Listing 3, with a subset of metadata file reported in Listing 4. As an example use case of interest for biologists, G4 predicted regions could be used to identify genes whose expression can be potentially inhibited by folding a preceding G4 structure. Once obtained this genes (together with their locations), we could find a correlation with known disease and such a search could be easily done genome wide. Another interesting use case of interest by pharmacologists and physicians is related to the cylindincal shape of a G4 structure, a potential binding site for drugs which have the cell as a target.

```
DATA = SELECT (cell == K562 AND dataType == ChipSeq) S1;
GENES = SELECT (annotation_type == 'genes') S1_ANNOTATION;
G4 = SELECT (annotation_type == 'G4') S1_ANNOTATION;
G4VALID = PROJECT (accuracy > 0.5) G4;
G4GENES = JOIN (FIRST AFTER DOWNSTREAM_DISTANCE 0,
                RIGHT_DISTINCT) G4VALID, GENES;
```

Listing 5: GMQL query example: Select the genes which can be inhibited by inducing a G-Quadruplex folded region; consider only G4s having prediction accuracy greater than 0.5

In Listing 5 a typical query in GMQL to the GenData2020 system is reported, where a list of potentially inhibited genes are searched in the genome database by selecting all the genes being in positions right next the predicted G4 regions.

References

[1] Tradigo, G., Mannella, L., Veltri, P., Assessment of G-quadruplex prediction tools, *In Computer-Based Medical Systems (CBMS)*, 243-246, 2014

[2] C.X. Xu, Y.X. Zheng, X.H. Zheng, Q. Hu, Y. Zhao, L.N. Ji, Z.W. Mao, V-Shaped Dinuclear Pt(II) Complexes: Selective Interaction with Human Telomeric G-quadruplex and Significant Inhibition towards Telomerase, *Scientific Reports*, 3, 2060, 2013

[3] Q. Li, J.F. Xiang, Q.F. Yang, H.X. Sun, A.J. Guan, Y.L. Tang, G4LDB a database for discovering and studying G-quadruplex ligands, *Nucleic Acids Research*, 41, 1115-1123, 2012

[4] Huppert, J.L, Structure, location and interactions of G-quadruplexes, *FEBS Journal*, 277, 3452-3458, 2010

[5] S. Neidle, S. Balasubramanian, Quadruplex Nucleic Acids, *RSC Publishing*, 2006

[6] R.Z. Cer, D.E. Donohue, U.S. Mudunuri, N.A. Temiz, M.A. Loss, N.J. Starner, G.N. Halusa, N. Volfovsky, M. Yi, B.T. Luke, A. Bacolla, J.R. Collins, R.M. Stephens, Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools, *Nucleic Acids Research*, 41, 94-100, 2013

[7] V.K. Yadav, J.K. Abraham, P. Mani, R. Kulshrestha, S. Chowdhury, QuadBase: genome-wide database of G4 DNA-occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes, *Nucleic Acids Research*, 36, 381-385, 2008

[8] R. Zhang, Y. Lin, C.T. Zhang, Greglist: a database listing potential G-quadruplex regulated genes, *Nucleic Acids Research*, 36, 372-376, 2008

[9] O. Kikin, L. DAntonio, P. S Bagga, QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences, *Nucleic Acids Research*, 2006, 34, 676-682, 2006

[10] L. DAntonio, P. Bagga, Computational Methods for Predicting Intramolecular G-quadruplexes in Nucleotide Sequences, *Computational Systems Bioinformatics Conference*, 2004

[11] J. Eddy, N. Maizels, Gene function correlates with potential for G4 DNA formation in the human genome, *Nucleic Acids Research*, 34, 3887-3896, 2006

[12] R.Z. Cer, K.H. Bruce, D.E. Donohue, N.A. Temiz, U.S. Mudunuri, M. Yi, N. Volfovsky, A. Bacolla, B.T. Luke, J.R. Collins, R.M. Stephens, Searching for non-B DNA-forming motifs using nBMST (non-B DNA Motif Search Tool), *Current Protocols in Human Genetics*, 18, 2012

[13] J.L. Huppert, Hunting G-quadruplexes, *Biochimie*, 90:8, 1140-1148, 2008

[14] Mirabello, C., Tradigo, G., Pollastri, G., Distill: protein structure prediction by Machine Learning, *Poster in IX Ed. of CASP, Critical Assessment of Techniques for Protein Structure Prediction*, 2012

[15] Tradigo, G., Greco, S., Veltri, P., G4Predictor website, *http://g4predictor.appspot.com/*, 2015