Software Requirements Specification

QGRS MAPPER

GROUP IV

Submitted in partial fulfilment
Of the requirements of
CS 258
Software Engineering

Bhor Verma	Keshav Goyal	Vinit Shah
150001005	150001014	150001029

1. Introduction

1.1. Purpose

The purpose of this document is to present a detailed description of the QGRS Mapper , a web-based server for predicting G-quadruplexes in nucleotide sequences. It will explain the purpose and features of the tool, the interfaces on which the tool which operate, what the system will do, the constraints under which it must operate.

1.2. Scope

This software system will be a Web App for all the users of interest and will be globally available on the internet . This system will be designed with an interactive interface which will be easy to use.
More specifically, this system is designed to view the complex nucleotide sequences in a very systematic manner with all the details.

1.3 Definitions, Acronyms, and Abbreviations

DNA is a molecule that carries the genetic instructions used in the growth, development,
functioning and reproduction of all known living organisms and many viruses.Each nucleotide is composed of one of four nitrogen-containing nucleobases either cytosine (C), guanine (G), adenine (A), or thymine (T) and a sugar called deoxyribose and a phosphate group.
G-quadruplexes are are tertiary structures formed in nucleic acids by sequences that are rich in guanine.Four guanine bases come together to form a square planar structure with metallic ions at centers.
QGRS is a **Q**uadruplex forming **G**-**R**ich **S**equences.
Nucleotides are organic molecules that serve as the monomers, or subunits, of nucleic acids like DNA(deoxyribonucleic acid) and RNA (ribonucleic acid).The building blocks of nucleic acids, nucleotides are composed of a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group. Thus a nucleoside plus a phosphate group yields a nucleotide.

The **binding constant**, or **association constant (Ka)**, is a special case of the equilibrium constant *K*, and is the inverse of the dissociation constant. It is associated with the binding and unbinding reaction of receptor (R) and ligand (L) molecules.
A **Hoogsteen base pair** is a variation of base-pairing in nucleic acids such as the A•T pair. In this manner, two nucleobases, one on each strand, can be held together by hydrogen bonds in the major groove.
A **chromosome** is a packaged and organized structure containing most of the DNA of a living organism. Most eukaryotic cells have a set of chromosomes (46 in humans) with the genetic material spread among them.
A **glycosidic bond** or **glycosidic linkage** is a type of covalent bond that joins a carbohydrate (sugar) molecule to another group, which may or may not be another carbohydrate.

1.4 REFERENCES

1. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences -> https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538864/

2. 34. Huppert J.L., Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33:2908–2916.
3. 35. Todd A.K., Johnston M., Neidle S. Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res. 2005;33:2901–2907.
4. 39. Hazel P., Huppert J., Balasubramanian S., Neidle S. Loop-length-dependent folding of G-quadruplexes. J. Am. Chem. Soc. 2004;126:16405–16415.
5. 40. Petraccone L., Erra E., Duro I., Esposito V., Randazzo A., Mayol L., Mattia C.A., Barone G., Giancola C. Relative stability of quadruplexes containing different number of G-tetrads. Nucleosides Nucleotides Nucleic Acids. 2005;24:757–760.

1.5 Overview

- **Existing system**
  - Browse a nucleotide sequence from database
  - Quick search
  - Advanced search
  - G4-predictor tool

- **Our plan**
  - Improve the search algorithm of QGRS mapper
  - Addition of Graphical View
  - Upgradation of User Interface
  - Data View with overlaps
  - Sequence view
  - Added security features
  - Addition of cA score
  - Option for minimum G- group size
  - Option for selecting
  - Option for selecting maximum length of sequence
  - Option for selecting loop size of sequence

2. General Description

The main goal of the QGRS Mapper program is to predict the presence of QGRS in nucleotide entries. These putative G-quadruplexes are identified using the following motif.

$$G_xN_{y1}G_xN_{y2}G_xN_{y3}G_x$$

Here x = number of guanine tetrads in the G-quadruplex and $y1$,$y2$,$y3$ = length of gaps (i.e. the length of the loops connecting the guanine tetrads). The motif consists of four equal length sets of guanines (which we call G-groups), separated by arbitrary nucleotide sequences, with the following restrictions.

- The sequence must contain at least two tetrads (i.e. $x \geq 2$). Although structures with three or more G-tetrads are considered to be more stable, many nucleotide sequences are known to form quadruplexes with two G-tetrads . QGRS Mapper is meant to be a flexible and comprehensive tool for investigating G-quadruplexes; hence it considers sequences with two tetrads.
- By default, only QGRS of maximum length of 30 bases are considered. However, the program gives the user the option to search for sequences up to 45 bases. This restriction on the length of the sequences being considered is in agreement with recent literature . The maximum length of 30 bases restricts G-groups to a maximum size of 6.
- The gaps or loops between the G-groups may be arbitrary in composition or length (within the overall restrictions on the length of QGRS). The program gives the user the option to search for QGRS having loops with a specified length range. The user can also specify a string that one or more loops of each QGRS must contain. This string can be given as a regular expression. For example, entering the regular expression 'T{3,5}' will search for QGRS having one or more loops that contain three to five consecutive T's.
- Also, at most one of the gaps is allowed to be of zero length

The first sequence has four tetrads and equal length gaps. This would seem to provide a G-quadruplex that is the most stable of the three sequences. The second sequence is notable for the significant differences in the size of its loops. The third sequence has two tetrads, even though three of the G-groups could have included another G (since all G-groups must be equal in size).

2.1 Software Perspective

The software is supposed to be an proprietary. The software tool mainly consists of two parts: the sequence analyzer and the database. The database contains the information about the various types of protein sequences and the sequence analyzer performs different type of operations on the input sequence and calculates a number of scores and graphs for analyzing the sequence. The input nucleotide sequence can have a maximum length of 10^5.

2.2 Software Functions

2.2.1 Design and implementation

QGRS Mapper is a web-based program, written in PHP, with Javascript being used for its graphics. The program takes a nucleotide sequence as provided by the user and analyzes it for the presence of putative G-quadruplexes.

2.2.2 Search and analysis

QGRS Mapper allows the user to search for putative G-quadruplexes in a variety of ways. It is possible to enter a nucleotide sequence in raw or FASTA format for analysis. The user can opt to change the maximum length of QGRS that will be searched for (the default maximum length being 30) and change the minimum sized G-group (which is two by default)

3. Specific Requirements

This section contains all of the functional and quality requirements of the system. It gives a detailed description of the system and all its features.

3.1 External interface Requirements

This section provides a detailed description of all inputs into and outputs from the system. It also gives a description of the hardware, software and communication interfaces and provides basic prototypes of the user interface.

3.1.1 User interfaces

The user will get option to enter the query nucleotide sequence either by:

1. By manually entry of the sequences.
2. By uploading the sequence file directly from the directory in FASTA file format or in text file format.
3. By defining the number of G that will appear together in the sequence and the number of the N so as to define the structure of the G-quadruplex.

After entering the sequence, the user will get the total number of overlapping QGRS sequences found in the input sequence. The user will get the option of viewing each QGRS sequence in a tabular form, in sequence form and also in a graphical view.
On clicking the tabular view, the user will see all the starting and ending positions of the found sequences. Along with the positions, it will also show the calculated cG, cA and cC score according to the algorithm.
On clicking the sequence view, the user will get to see the sequence view of all the QRRS sequence. User will get to see a manhattan of the position of the sequences v/s the cG score. The user will also get to see the zoomed view of the sequences along the found QGRS sequences being marked differently in another color. On zooming the sequence the user will get to see the exact sequence marked.
On clicking the graphical view, the user will get to see a graph of the QGRS sequences v/s their found positions. All the found will be both overlapping and overlapping, i.e. the user will get to see all two tables, one containing the overlapping nucleotide sequences and one containing the non overlapping nucleotide sequences.

3.1.2 Hardware Interfaces

Since the web portal have any designated hardware, it does not have any direct hardware interfaces. The GQRS is managed by the application of the tool by the algorithm and the hardware connection to the database server is managed by the underlying operating system on the web tool.

3.1.3 Software Interfaces

The communication between the user and the tool consists of reading operations. The algorithm then works on the input sequence and then calculates the required cG, cC and cA scores. This information is then stored in the web tool and the user is given different options to analyze the sequence like the number of QGRS sequences between two positions, the highest and the lowest cG scores, number of QGRS sequences having a cG/cC score greater than a threshold value, the loop length, etc.

3.2 Functional Requirements

3.2.1 Quick search option

In order to search database rapidly rather than constructing the proteins structure, there is Quick search option in the database. This search option facilitates the user to search database directly by protein or target name, target sequence, protein sequence, UniProt-ID, UniProt entry name, author's name, technique used in the study, PMID etc. For example, if user gives FMRP as query string, this search option gives the output that has FMRP as interacting protein.

3.2.2 Advanced search option

There is an advanced search option to facilitate users to search database specifically and efficiently on the basis of user's choice. In this search option there are 10 different types of search criteria such as G4IPDB interaction ID, DNA/RNA target name, DNA/RNA target sequence, Interacting protein name, UniProt-ID, UniProt entry name, protein coding gene name, gene synonyms, PMID and author's name. User can simultaneously select the range of search criteria and get the search result in more specific manner.

3.2.3 G4 sequence predictor tool

From last few decades, guanine rich sequences has gravitated the attention of scientific community because of their regulatory role in various biological processes and as a potent therapeutic target. Therefore, building of efficient computational tool to mine the putative G-quadruplex forming sequences in the genome is highly important. G4IPDB provides a web-based tool that predicts the putative G-quadruplex forming sequences based on the manually designed algorithm. G4-predictor tool is capable of predicting the putative G-quadruplex forming sequence simultaneously in both sense and antisense strands. It also shows the output of the start and end positions of each putative G-quadruplex forming motif and provides total number of putative G-quadruplex motif in the queried sequence. G4 predictor tool is extensively user friendly that allows user to either enter the sequence manually or browse the sequence containing file from the local disk. It is efficient to perform analysis of very large sequences on any genome size ranging from bacteria to mammalian genome. The prediction for putative G-quadruplex forming sequences is based on pattern matching of G{Y1}[X]{Y2}G{Y1}[X]{Y2}G{Y1}[X]{Y2}G{Y1} motif. In this motif G represented the Guanine nucleotide and X represented any nucleotide including Adenine (A), Guanine (G), Cytosine(C), Thymine (T) and Uridine (U) nucleotide. The length of guanine tracts (Y1) varies from 2 to 7 in number and length of loop (Y2) varies with minimum of 1 and maximum of 7 nucleotides. The stability of predicted G-quadruplex structure will depend upon the number of guanine tracts, length of the internal loop as well as on the number of tandem repeats of the motif sequence. The default maximum length of putative G-quadruplex forming sequence is 49 bases.

3.2.4 QGRS Mapper output

After the analysis of overlaps is completed, QGRS Mapper displays a summary of its findings, in the Gene View. This summary includes basic gene information such as the gene ID, gene symbol, gene name, a link to the NCBI entry, organism name, chromosome number and number of products and poly(A) signals. Information is also given for each product, such as the number of exons and introns, number of QGRS (non-overlapping and overlapping), number of QGRS found near RNA processing sites, and a visual map of the product.

At this stage in the analysis the user can choose among three further displays: ‘Data View’, ‘Data View (with overlaps)’ and ‘Graphics View’. This can be done for the entire gene or for any particular product.

In the Data View, a table is displayed showing information for each of the set of non-overlapping QGRS. This table displays the position of the QGRS, which exon/intron it appears in, its distance from 3′ and 5′ splice sites, the QGRS sequence (with each G-group underlined) and the corresponding G-score. Similar display is also shown for each QGRS mapped to poly(A) region in the product. If the user requests the Data View for the entire gene, then the QGRS information is shown for each product. The 'Data View (with overlaps)' gives the same information but shows the locations of all QGRS.

The user can also choose the Graphics View to give a visual display of the location of QGRS. This allows the user to see the location of QGRS relative to exons and introns (if that information is available). The Graphics View has the following components.

- A graphic display of the entire gene (showing the location of the exons). This display includes a sliding window that can be used to focus on any particular segment of the gene. This window may be dragged to the left or right to change position within the gene.
- A magnified view of the fragment of the gene within the sliding window.
- A graph showing the location of QGRS within the fragment, with each QGRS being displayed by a bar whose height represents its G-score.
- A vertical slider that allows the user to change the size of the window. This allows the user to zoom in or out on any part of the gene. The sliding window on the gene expands or contracts as one zooms in or out. It is possible to see the nucleotide sequence of the product at maximum zoom levels.

The Graphics View for the entire gene shows the G-score graph together with an exon/intron map for each product. This allows the user to visually compare the location of QGRS for each product relative to that of splice sites.

3.3 Non Functional Requirements

3.3.1 Performance

The system must be interactive and the delays involved must be less .So in every action-response of the system, there must be no immediate delays. In case of opening windows forms, of popping error messages and saving the settings or sessions there should be minimum delay. In case of opening databases, analyzing the QGRS sequence and evaluation of the cG, cC and cA scores, there should be a minimum delay. Also when connecting to the server the delay should be less.

3.3.2 Reliability

As the system provide the right tools for analyzing, sequence analyzing must be made sure that the user is reliable in its calculations

3.3.3 Security

The main security concern is for protein database. Hence proper validations should be used to avoid hacking. The input sequence should be first validated and is trimmed of the special characters, so as to avoid any injections on the page. Hence, security is provided from unwanted use of protein database.

3.3.4 Availability

The web tool would be then made available over the World Wide Web.

3.4 Design Constraints

The graph which shows the QGRS sequence position v/s the cG score and the graph which shows the input nucleotide sequence v/s the positions where the QGRS sequences are found can be enlarged or diminished.After zooming in to a specific level the user would be able to see the user input sequence along with its highlighted QGRS sequence and cG score.

APPENDIX

Appendix 1

G-quadruplex

In molecular biology, **G-quadruplexes** (also known as **G_4-DNA**) are tertiary structures formed in nucleic acids by sequences that are rich in guanine. Four guanine bases can associate through Hoogsteen hydrogen bonding to form a square planar structure called a guanine tetrad, and two or more guanine tetrads can stack on top of each other to form a G-quadruplex. The quadruplex structure is further stabilized by the presence of a cation, especially potassium, which sits in a central channel between each pair of tetrads.[1] They can be formed of DNA, RNA, LNA, and PNA, and may be intramolecular, bimolecular, or tetramolecular.[2] Depending on the direction of the strands or parts of a strand that form the tetrads, structures may be described as parallel or antiparallel.

The quadruplex structures formed by guanine-rich nucleic acid sequences have received significant attention recently because of increasing evidence for their role in important biological processes and as therapeutic targets (1–5). The G-quadruplex structure, also known as a G-quartet, is formed by repeated folding of either the single polynucleotide molecule or by association of two or four molecules. The structure consists of stacked G-tetrads, which are square co-planar arrays of four guanine bases each (6). G-quadruplex is stabilized with cyclic Hoogsteen hydrogen bonding between the four guanines within each tetrad. The present work focuses only on the unimolecular quadruplexes, since it is more likely to be encountered in physiological conditions .

Guanine-rich sequences capable of forming G-quadruplexes are found in telomeres, promoter regions, transcribed and other biologically important regions of the mammalian genomes. G-quadruplex DNA has been suggested to regulate DNA replication in retinoblastoma susceptibility gene (Rb) region . This structure may control cellular proliferation at telomeric level and by transcriptional regulation of oncogenes like *c-myc* and *c-kit* . Formation of G-quadruplex seems to be regulated through interactions with cellular proteins. While some proteins help stabilize the structure , others are known to resolve it . Proteins and chemicals that stabilize the G-quadruplex structure can inhibit telomerase action and, therefore, are being evaluated as anticancer therapeutic agents . Chemical compounds that inhibit G-

quadruplex helicase activity may also be capable of regulating cellular proliferation . G-quadruplexes are also being eyed as potential antimicrobial agents due to their ability to transport monovalent anions .

APPENDIX: 2

G4 Prediction Score

In combination of the mining and prediction of putative G-quadruplex motif, G4 predictor tool also calculates the 'cG', 'cC', and 'cG/cC' scores based on the previous study of new scoring function for G-quadruplex motif. Briefly, the cG score calculation is based on the following equation and applied for each predicted substring that has the length of n:

$$cG(s) = \sum_{i=1}^{n} (|Gs(i)| \times 10 \times i) \qquad (1)$$

In this equation (1) a value of 10 is assigned to the each G, a value of 20 assigned for each paired GG and a value of 30 assigned for each triplet GGG and so on. The cC score calculation is also based on the similar equation only difference is that the cytosine nucleotide is used in place of guanine nucleotide. The cG/cC score is based on the ratio of both cG and cC scores. Nucleic acid G-quadruplex structure (G4) Interacting Proteins DataBase (G4IPDB) is an important database that contains detailed information about proteins interacting with nucleic acids that forms G-quadruplex structures. G4IPDB is the first database that provides comprehensive information about this interaction at a single platform. This database contains more than 200 entries with details of interaction such as interacting protein name and their synonyms, their UniProt-ID, source organism, target name and its sequences, ΔTm, binding/dissociation constants, protein gene name, protein FASTA sequence, interacting residue in protein, related PDB entries, interaction ID, PMID, author's name and techniques that were used to detect their interactions.G4IPDB also provides an efficient web-based "G-quadruplex predictor tool" that searches putative G-quadruplex forming sequences simultaneously in both sense and antisense strands of the query nucleotide sequence and provides the predicted G score. Studying the interaction between proteins and nucleic acids forming G-quadruplex structures could be of therapeutic significance for various diseases including cancer and neurological

disease, therefore, having detail information about their interactions on a single platform would be helpful for the discovery and development of novel therapeutics.It generates putative **QGRS** sequences in user input nucleotide sequences.