Quantitative Structural Model for Prediction and Analysis of R-loop Forming Sequences in the Genomes

Piroon Jenjaroenpun

*Bioinformatics Institute, A*STAR, Singapore, 138671,* piroonj@bii.a-star.edu.sg

Thidathip Wongsurawat

*Bioinformatics Institute, A*STAR, Singapore, 138671,* piroonj@bii.a-star.edu.sg

Surya Pavan Yenamandra

*Bioinformatics Institute, A*STAR, Singapore, 138671,* piroonj@bii.a-star.edu.sg

Vladimir A. Kuznetsov

*Bioinformatics Institute, A*STAR, Singapore, 138671,* piroonj@bii.a-star.edu.sg

INTRODUCTION

The R-loop is a three-stranded nucleic acid structure that is co-transcriptionally formed RNA-DNA hybrid between a nascent guanine-rich RNA transcript segment and a DNA template whilst leaving the non-template DNA strand in a single-stranded conformation. This structure forms naturally during transcription and has been detected in a wide range of organisms from bacteria to humans. R-loops lie at the interface of multiple biological processes, including RNA transcription and processing, RNA, DNA and chromatin interactions, DNA damage, mutagenesis and cell proliferation and differentiation. Altering the R-loops balance can impair R-loop-mediated processes, resulting in mutagenesis and genome instability and possibly leading to various diseases. Targeting RNA-DNA hybrids in R-loops using small molecules has the potential to be clinically important; thus, these types of strategy are currently under development (1).

The systematic mapping and prediction of R-loops are key issues for the structural and functional characterisations of R-loops (2). In 2011, we published the 1-st quantitative structural model of RLFSs (3), whose parameters were specified based on a few in vitro and in vivo data publicly available at that time. Surprisingly, the model predicted that these three-stranded RNA and DNA hybrid structures could be formed in ~ 60% of human genes. RLFSs were preferentially predicted in thousands of functionally-important guanine-rich genic and inter-genic regions, in the 1st exons and the 1st introns, 3'UTRs, downstream

poli(A) tails, telomere ends, disease critical regions. These findings consist of the results of experimental studies. We also developed the R-loop prediction database, providing detail information about the sequence structures and locations of RLFSs and in genic regions of the human genome (http://rloop.bii.a-star.edu.sg/). Here, we extended our RLFS model and hypothesized that R-loops might be identified in many thousand genes and the genomes of different species, and might be hotspots in genes that have a predisposition to 'catastrophic' patho-biological events leading to many diseases.

METHODS

Generalized structural motif models of RLFSs

Here, we generalized our original RLFS model (4) and used that generalized model to develop a pipeline for predicting the structures, locations and functions of RLFSs on the genome scale. Our initial computational models of RLFS (4) has identified three structural features in DNA sequences, including a short G-cluster rich region responsible for initiating R-loop formation (R-loop initiation zone or RIZ), a structurally non-specified linker (linker) and a downstream region that is relatively long and has a high G-density R-loop elongation zone (or REZ).(Figure 1).The three zones (or sequence elements) constitute the RIZ-Linker-REZ configuration and form the basis of our computational RLFS prediction model. Such sequence elements and their configuration in the non-template DNA sequence have been proposed in(5) based on biochemical and molecular biology studies of the roles of the G-clusters and high G-density sequences in transcriptional R-loop formation. Using the characteristics of empirical R-loop sequence models (5), the computational model (4) predicts the locations of RLFS in the genes of human genome.

We improved the sensitivity of RIZ detection by incorporating an additional sequence composition for RIZ into our original RLFS predictive model (4). In additionally to the previous model, the generalised model can also use two linked G-clusters as a RIZ. The number of contiguous-Gs increased to four guanines, instead of the maximum of three guanines per G-cluster in the basic model.

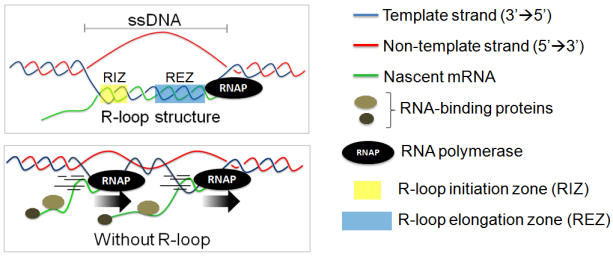

Figure 1. Structural model of R-loop (top) and transcriptional model without R-loop formation (bottom).

The main aim of the QmRLFS-finder program is to predict the presence and location of the RLFSs in nucleotide sequences (intragenic, extra-genic or artificial nucleic acid sequences) . The RLFSs are identified using a simple pattern-based rule by matching sequences of the form:

$$\underbrace{G_{n1}N_xG_{n1}N_xG_{n1}}_{RIZ},\ \underbrace{N_y}_{linker},\ \underbrace{N_z}_{REZ}\ \begin{cases} RIZ \geq 50\%G \\ REZ \geq 40\%G \end{cases}, \qquad \text{(m1)}$$

$$\underbrace{G_{n2}N_xG_{n2}}_{RIZ},\ \underbrace{N_y}_{linker},\ \underbrace{N_z}_{REZ}\ \begin{cases} RIZ \geq 50\%G \\ REZ \geq 40\%G \end{cases}, \qquad \text{(m2)}$$

where G_{n1} and G_{n2} are the guanine cores that can occur with different numbers of G-residues ($n1 \geq 3$ and $n2 \geq 4$, respectively) in RIZ sequence feature. The RIZ in (m1) contains 3 G-clusters, and the RIZ in (m2) contains 2 G-clusters. The symbol N_x denotes a sequence with non-specified composition of nucleotides, where the number of nucleotides (x) $1 \leq x \leq 10$. Additionally, RIZ must contain at least 50% guanine content in the both models. For the Linker sequence feature, the symbol N_y denotes an arbitrary nucleotide composition sequence number (y) $0 \leq y \leq 50$ which do not affect R-loop extension (3). For the REZ sequence feature, the symbol N_z denotes an arbitrary nucleotide composition sequence number (z) $100 \leq z \leq 2000$. REZ requires at least 40% guanine content which is considered as important feature to maintain R-loops formation (3,4). By our model, the REZ region can also include the G-cluster regions.

RESULTS

Our generalized quantitative structure RLFS predictive model (m1) and (m2) is not *a priory* limited by any other sequence composition constrains or preselected regulatory signals, including CpG islands, repeats, gene loci, genome architectures or genome.

Our comparison of the 22 QmRLFS prediction models with the 22 experimentally detected loci forming and not forming R-loops (of 17 publicly available loci and 5 loci detected in cancer cells by our group using DRIP-qPCR with S9.6 Ab) demonstrated consistency in the 21 cases. Table 1 demonstrates the statistics of predicted RLFSs and RLFS clusters in human and 4 model organism genomes. This table shows that the most genes in the mammalians and chicken are highly populated by RLFS and RLFS clusters. These sequences often found in promoter regions of the RLFS-positive genes. The RLFS loci often showed evolution conservation across the mammalians. Co-localization analysis showed frequent colocation of RLFS with G-quadruplexes, CpG islands and the genomic regions functionally related to cancer and neurodegeneration diseases.

Table 1. The number of predicted RLFSs and RLFS clusters in the human and four model organism genomes

Organism	Number of RLFSs	Number of RLFS clusters	Total number of all genes	Number and % genes containing RLFSs
human	664,773	110,334	25,844	19,573 (75.74%)
mouse	569,574	106,448	24,017	17,415 (72.51%)
rat	454,018	87,476	17,168	12,565 (73.19%)
chicken	169,184	31,409	6,289	4,314 (68.60%)
fruit fly	5,910	1,069	16,612	1,674 (10.08%)

CONCLUSION

QmRLFS model demonstrates highly-accurate predictions of the detected RLFSs, proposing new perspective to further discoveries in the R-loop biology, evolution, disease involving, biotechnology and molecular therapy.

FUNDING & ACKNOWLEDGEMENTS

Biomedical Research Council of Agency for Science, Technology and Research (A*STAR), Singapore. The authors are grateful to Dr. Martin Lavin and Ms. Abrey Yeo for providing us with the S9.6 antibody._

1. Shaw, N.N., Xi, H. and Arya, D.P. (2008) *Bioorg Med Chem Lett*, **18**, 4142-4145.
2. Aguilera, A. and Garcia-Muse, T. (2012) R loops: from transcription byproducts to threats to genome stability. *Mol Cell*, **46**, 115-124..
3. Wongsurawat, T., Jenjaroenpun, P., Kwoh, C.K. and Kuznetsov, V. (2012) *Nucleic Acids Res*, **40**, e16.(on line publ. 2011)
4. Roy, D. and Lieber, M.R. (2009) *Mol Cell Biol*, **29**, 3124-3133.