Fast identification of statistically significant lncRNA-RNA interactions in a transcriptome-wide search

Ivan Antonov
Institute of Bioengineering, Research Centre of Biotechnology, RAS, Moscow, Russia, ivan.antonov@gatech.edu

Andrey Marakhonov
Federal state scientific budgetary Institution "Research Centre for Medical Genetics", Moscow, Russia, marakhonov@gmail.com

Maria Zamkova
Russian N.N.Blokhin Cancer Research Center, Moscow, Russia, zamkovam@gmail.com

Mikhail Skoblov
Federal state scientific budgetary Institution "Research Centre for Medical Genetics", Moscow, Russia, mskoblov@gmail.com

Yulia Medvedeva
Institute of Bioengineering, Research Centre of Biotechnology, RAS, Moscow, Russia, ju.medvedeva@gmail.com

Long noncoding RNAs (lncRNAs) are a large and diverse class of transcribed RNA molecules with a length of more than 200 nucleotides that do not encode proteins. The discovery of thousands of lncRNAs in mammals raised a question about their functionality. Due to low conservation and functional diversity the role and/or molecular mechanism of only few hundred lncRNAs have been determined by the date. Particularly, it has been shown that some of them function post-transcriptionally via formation of inter-molecular RNA-RNA duplexes. Such RNAs are known as natural antisense transcripts (NATs).

Several experimental methods have recently been developed to identify intermolecular RNA-RNA duplexes on a large scale [1-3]. Nevertheless, due to the limited availability of the experimental data there is still a need for a computational prediction of possible antisense partners of a given lncRNA.

Existing computational tools [4-6] compute the free energy of the inter-molecular duplexes

(ΔG) to predict the RNA-RNA interactions. Although these algorithms are effective in working with relatively short RNAs (such as bacterial sRNAs) on a small scale, our analysis indicates that they have poor performance when applied to lncRNAs for transcriptome-wide searches. The two main problems include the large execution time as well as the absence of the statistics-based measure of the interaction strength.

To address these challenges we develop a new computational pipeline ASSA ("AntiSense Search Approach") that combines sequence alignment and thermodynamics tools for efficient prediction of RNA-RNA interactions between long transcripts. It reduces the running time by fast identification of the putative antisense sites using the sequence alignment tool LASTAL. Next, the detected sites along with the flanking sequences are extracted from the transcripts and the interaction energy of every site (“ΔG-site”) is calculated using a thermodynamics-based tool. It should be noted that application of the thermodynamics-based tool to the putative sites instead of the full-length transcripts significantly speeds up the calculation of the interaction free energy .

Next, the sum of the free energies from all the putative sites between two RNAs (“sum-ΔG”) is computed in order to measure the interaction strength. Clearly, the value of the sum energy depends on several factors including the transcript lengths (longer transcripts produce more putative sites), GC-content (G::C base-pairing is stronger than the A::T) as well as the parameters of the LASTAL search and the length of the flanking sequences. To compute the statistical significance of the observed sum energies a distribution of these values for random sequences is needed. Thus, we develop a model to obtain the parameters of the sum score distribution with respect to the properties of the input sequences as well as the program parameters. In short, ASSA is applied to a number of random sequences with various lengths and GC-contents and for particular fixed values of the sequence properties and the ASSA parameters (LASTAL score threshold and the flank length) the empirical distribution of the observed sum-ΔG values is generated. The mean and variance is computed for every distribution. Finally, two polynomial regression models are fitted to the observed dependences of the distribution mean and variance on the different sequence features and

ASSA settings. Using simulated NATs we optimize the default values for the LASTAL threshold and the flank length. The obtained models are used to estimate the statistical significance (p-value) of the interaction between two transcripts taking into account their lengths and GC content.

ASSA outperforms four other tools in speed and accuracy. Moreover, we demonstrate that ASSA can be used to elucidate the lncRNA mechanisms. Particularly, our results for the lncRNA HOTAIR support the model of its binding to the chromatin through direct hybridization with the nascent transcripts. We believe that ASSA will be a useful tool to both bioinformatics and wet-lab researches.

The authors would like to thank Andrey Leontovich and Alexandra Filatova for useful discussions; Yuriy Shkandybin for technical support and the Dynasty foundation for the financial support (Postdoctoral Fellowship No DP-B-26/14 to IA).

1. Z. Lu et al (2016) RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, **165(5)**: 1267–79.

2. J. Aw et al (2016) In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol Cell*, **62(4)**: 603–17.

3. E. Sharma et al (2016) Global Mapping of Human RNA-RNA Interactions. *Mol Cell*, **62(4)**: 618–26.

4. L. DiChiacchio et al (2016) AccessFold: predicting RNA-RNA interactions with consideration for competing self-structure. *Bioinformatics*, **32(7)**: 1033-9.

5. J. Li et al (2015) LncTar: a tool for predicting the RNA targets of long noncoding RNAs. *Brief Bioinform*, **16(5)**: 806-12.

6. H. Tafer et al (2011) Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics*, **27(14)**: 1934-40.