Antisense interactions of long noncoding RNAs in human cells

Ivan Antonov

Research Centre for Medical Genetics, Moscow, Russia

ivan.antonov@gatech.edu

Mikhail Skoblov

Research Centre for Medical Genetics, Moscow, Russia

Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

mskoblov@generesearch.ru

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 existence of individual lncRNAs has been known for several decades. Little was known about the diversity of lncRNAs in mammalian genomes until the advent of technologies that were capable of unbiased high throughput sequencing of all the expressed transcripts in cells (RNA-seq). The discovery of thousands of lncRNAs in human and mouse transcriptomes raised a question about their functionality.

By the date the role of only few hundred lncRNAs has been determined. Some of them function via formation of inter-molecular RNA-RNA duplexes. Such RNAs are known as natural antisense transcripts (NATs). Moreover, it is known that lncRNAs can regulate expression of protein coding genes at the post-transcriptional level (regulation of mRNA stability or translation). In this work we predicted the cases of gene regulation via formation of lncRNA-mRNA duplexes.

Several tools aimed to predict RNA-RNA interactions by minimizing the free energy of the overall structure have been previously developed. The execution time of these algorithms exponentially depends on the total length of the input RNA sequences. This can be suitable to work with relatively short RNAs (such as miRNAs), but the prediction of lncRNA-mRNA interactions becomes computationally expensive. For example, *bifold* [1] takes up to 20 hours to predict interaction between two sequences of length 3000 nucleotides each. This makes it impossible to directly use these tools in transcriptome-wide searches that require thousands of comparisons. Here we present a computational pipeline for transcriptome-wide prediction of regulatory lncRNA-mRNA interactions. Using this pipeline we tried to find specific lncRNAs whose function is to regulate expression of the target genes (NATs).

The ASSA pipeline

We started the current project with development of a new bioinformatics pipeline, called ASSA (AntiSense Search Approach) for initial prediction and filtering of possible antisense pairs prior to applying a time-consuming thermodynamic based method *bifold*.

On the first step of the pipeline, the similarity search method BLASTn is used for fast screening of a large sequence database (such as Refseq). During our analysis of BLASTn hits we noticed that transcripts with Alu-repeats have significantly more antisense partners (from 1000 to 1500 more predictions) than the transcripts without Alu. It should be noted that even though some Alu-based RNA-RNA interactions regulate gene expression [2], they are not gene specific, i.e. any Alu-containing transcript can potentially interact with any other transcript that has repeat in the antisense direction. To focus on gene specific regulation and to decrease the number of BLASTn hits, Alu-repeats were masked in all transcripts prior to BLASTn search.

Next, all the predicted interactions are subjected to filtering process. Three filters (E-value < 10^{-3}, at least one BLASTn site ≥ 20 nt and interaction between all alternative isoforms) reduce the initial number of BLASTn hits up to 10 times. The obtained lncRNA-mRNA pairs are submitted to the thermodynamics based tool *bifold*. It takes two RNA sequences as input and predicts a secondary structure by minimizing the free energy allowing inter-molecular interactions (antisense duplexes).

For further experimental evaluation of the pipeline predictions we considered genes expressed in the HEK293 human cell line. ASSA was applied to the 71 expressed lncRNAs to search for possible antisense targets among the 7600 expressed protein-coding genes. Overall, there were 427 lncRNA-mRNA gene pairs with the total duplex length predicted by *bifold* ≥ 100 bp. The majority of the interactions were *trans*-NATs, i.e. the corresponding lncRNA and protein coding genes are located in different genomic loci.

It should be noted that we observed little consistency between the regions found by BLASTn and the *bifold* duplexes. Moreover, the total length of the antisense sites predicted by *bifold* was often larger. The reason for this is that with our settings BLASTn predicts an antisense site in place of a 12 nt continuous region with perfect complementarity (seed length = 12). Thus, some short antisense regions or duplexes with mismatches would not be detected. From another hand, a stable secondary structure element on one of the RNAs can prevent inter-molecular duplex formation even though an antisense region was found by BLASTn.

Selecting regulatory antisense interactions

Correlation of the expression levels between two genes in a number of tissues and cell lines may indicate their common or mutual regulation. This approach was previously used to search for regulatory NATs [3-5]. We computed the correlation of expression profiles between the lncRNA and the protein coding gene from each of the 427 predicted antisense interactions based on the data for 889 human tissues and cell lines available at the FANTOM5 database [6]. Significant correlation (Pearson's p-value < 0.001) was observed for 252 gene pairs: 159 had positive and 93 had negative correlation coefficients.

Sequence conservation frequently corresponds to underlying biological function. We compared the coordinates of the conserved elements (the regions conserved across vertebrates [7]) with the locations of the predicted RNA-RNA duplexes and selected the interactions where an overlap was found. To test the importance of such observations we collected a set of 41 human overlapping gene pairs known from literature where the common regions have biological functions. This set included at least 9 cases of posttranscriptional regulation via RNA-RNA interaction (*cis*-NATs). In total, conserved elements in overlapping regions were found in 38 among 41 gene pairs, including 8 (out of 9) gene pairs with RNA-RNA interactions. In all four published cases of regulatory lncRNA-mRNA *trans*-NATs [8-10] the lncRNA binding sites overlap at least one of the conserved mRNA regions.

The combination of the expression correlation and the conservation analyses allowed us to select several lncRNA-mRNA interactions for experimental validation of the predicted posttranscriptional regulation.

The current version of the ASSA web server and the database are available at: http://assa.generesearch.ru/

The authors would like to thank Andrey Marakhonov whose original ideas were implemented in the project; Andrey Leontovich, Maria Zamkova, Alexandra Filatova for useful discussions; Yuriy Shkandybin for technical support and the Dynasty foundation for the financial support.

1. Mathews, D.H., et al. (1999) Predicting oligonucleotide affinity to nucleic acid targets. *RNA*. **5**: 1458-69.
2. Gong, C. and L.E. Maquat (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*. **470**: 284-8.
3. Li, J.T., et al. (2008) Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. *Nucleic Acids Res*. **36**: 4833-44.
4. Mahmoudi, S., et al. (2009) Wrap53, a natural p53 antisense transcript required for p53 induction upon DNA damage. *Mol Cell*. **33**: 462-71.
5. Su, W.Y., et al. (2012) Bidirectional regulation between WDR83 and its natural antisense transcript DHPS in gastric cancer. *Cell Res*. **22**: 1374-89.
6. Forrest, A.R., et al. (2014) A promoter-level mammalian expression atlas. *Nature*. **507**: 462-70.
7. Siepel, A., et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. **15**: 1034-50.
8. Hu, G., Z. Lou, and M. Gupta (2014) The long non-coding RNA GAS5 cooperates with the eukaryotic translation initiation factor 4E to regulate c-Myc translation. *PLoS One*. **9**: e107016.
9. Yoon, J.H., et al. (2012) LincRNA-p21 suppresses target mRNA translation. *Mol Cell*. **47**: 648-55.
10. Yuan, J.H., et al. (2014) A long noncoding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell*. **25**: 666-81.