At least 6% of conserved miRNAs` sites are misaligned.

Prosvirov K.A., Mironov A.A., Soldatov R.A.

Moscow State University Faculty of Bioengineering and Bioinformatics,

Moscow, Leninskie gory, b. 73, prosvirov.k@gmail.com

MicroRNAs are ~22 nt endogenous non-coding RNAs. MicroRNAs direct post-transcriptional repression of mRNAs through complementary binding predominantly to 3`UTR. Complementarity between 2-7th nucleotides is required and is known as microRNA seed region. Additional complementarity of 1st and 8th nt are widespread and enchances interactions. MicroRNA binding sites are classified in four major types: 8-mer, 7mer-A1, 7mer-m8 and 6mer with the number indicating length of binding stretch[1].

Determination of microRNA targets is crucial for understanding of the regulation of gene expression. But it is still a complicated bioinformatic problem, because the main properties that distinguish *bona fide* miRNA targets are poorly undestood. As a consequence, results of modern algorithms have substantial differences with experiments and each others[2]. To reduce the number of false-positive results and increase the accuracy of predictions, comparative genomics is widely used as the powerful tool for detection of conserved functional elements. Predictions of comparative genomics are based on multiple sequences alignments (MSA), which is prone to errors.

Overall alignment accuracy depends on the species divergence[3], for example human-mouse alignment contains ~15% of misaligned nucleotides[4]. While close species have no computational power for comparative analysis, divereged species face to the problem of incorrect MSA.

MSA accuracy doesn`t bypass the detection of conserved microRNA targets. Here we estimate the number of conserved microRNA sites (8mer, 7mer-A1, 7mer-m8)

missed due to incorrect MSA, and its dependence on evolutionary rate and species divergence.

We extend definition of the conserved microRNA binding site to account for local misalignments: the site of reference species is conserved in MSA if it can be found in window of size L in those species (Fig. 1). For simplicity, we call this L-conserved.

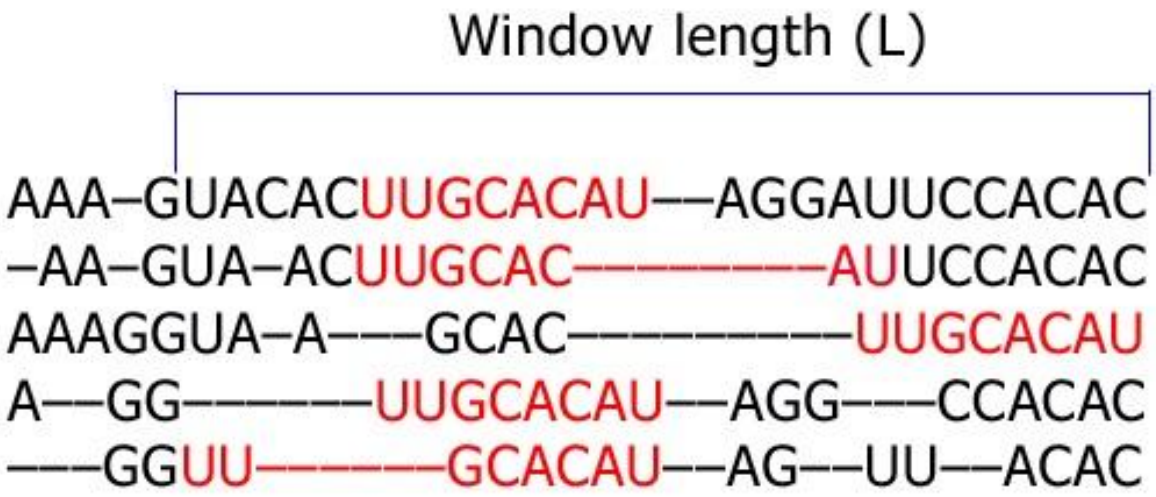

Fig.1. Definition of L-conserved site, red text is microRNA site.

We observe that at least 6% of conserved microRNA sites of all types is missed due to misalignment. The saturation of missed sites is reached at the window of L=25 nt, that indicate of local misaligment. Moreover, missed sites are genuine, because L-conservative approach shows similar sensitivity. The fraction of missed sites depends on evolutionary rate of 3'UTRs and overall divergence between species, Finally, we suggest that "relaxed" definition of conserved sites can substantially account for alignment errors, while retain computational power of method.

As a conclusion, we show that accuracy of MSA substantially affects conservative microRNAs sites with dependence on evolutionary rate and divergence time. We observed the significant increase in the number of detected sites of all types, also estimate the most suitable frame for detection of sites.

The work was supported by the Russian Science Foundation (grant 14-14-00088).

1. Robin Friedman, David Bartel et al (2009) Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. 2009:92-105.

2. Witkos TM et al (2011) Practical Aspects of microRNA Target Prediction, Current Molecular Medicine 2011 Mar;11(2):93-109.

3. Daniel A Pollard et al (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments, BMC Bioinformatics 2006,7:376.

4. Gerton Lunter et al (2008) Uncertainty in homology inferences: Assessing and improving genomic sequence alignment, Genome Research 2008 Feb; 18(2):298-309.

5. Lewis BP et al (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, Cell 2005 Jan 14;120(1):15-20.