Machine-Learning pipeline to predict relevant RNA motifs that allow the physical association between lincRNAs and androgen receptor

Vinícius Mesel[1], Lucas Ferreira DaSilva[1,2], Felipe Beckedorff[1,3], João C. Setubal[2,3], Sergio Verjovski-Almeida[1,2,3]

[1]Instituto Butantan, São Paulo, Brazil; [2]Pós-graduação Interunidades em Bioinformática and [3]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil

Introduction: The Androgen Receptor (AR) transcription factor is an important protein that acts in the regulation of prostate cells normal function and also has a crucial role in cancer development and progression. Under androgen hormonal stimulation the AR can modify the transcriptional program of the cell and change its epigenetic landscape through a complex signaling network. A large amount of evidence suggests that long intergenic noncoding RNAs (lincRNAs) can contribute to cell transcriptional changes. Some lincRNAs act regulating genes in their genomic neighborhoods by recruiting specific proteins like transcription factors and histone modifying enzymes.

Objectives: To investigate the importance of lincRNAs and AR association in the transcriptional regulation of LNCaP (Androgen-Sensitive Human Prostate Adenocarcinoma) cells under hormonal stimulation, and to identify the motifs that enable lincRNAs-AR protein binding.

Methods: A high-throughput RNA immunoprecipitation sequencing (RIP-seq) approach with specific anti-AR antibody was used to generate a list of lincRNAs physically associated with AR. To eliminate experimental binding noise, we used a non-specific IgG-antibody as a negative control. To identify the motifs that enable lincRNAs-AR protein binding we applied machine-learning algorithms to our sequences dataset to extract relevant patterns in the lincRNAs. To run the machine-learning algorithm and capture the precise association motifs, we grouped the sequences into two different groups: one of AR-associated lincRNAs (ARA-lincRNAs) and the other of IgG-associated lincRNAs (IgGA-lincRNAs). To separate the ARA-lincRNAs and IgGA-lincRNAs using the sequences features, we have split each sequence into *kmers* and we have extracted the sequences of all possible *kmers* of size *i (i = {5, 6, 7, 8, 9}).* The *kmers* profiles were loaded to a classification machine-learning algorithm.

Results and Discussion: The best performance (70 % success rate in a 10-fold cross-validation) was obtained by using *kmer* size i = 8. The most relevant 8-mer *kmers (z-score > 2)* used in classification were selected for subsequent analyses. GLAM2 (Gapped Local Alignment of Motifs tool) was used to reconstruct the conserved motifs of the original sequences. As GLAM2 input, we have used the most relevant *kmers* from the classification step. The motifs generated in GLAM2 were matched against the original lincRNA sequences using FIMO (Find Individual Motif Occurrences). The matched motifs were processed, generating a UCSC Genome Browser track for visualizing the specific position of motifs in lincRNAs regions.

Using this motifs extraction methodology we have been able to identify for the first time the RNA patterns that putatively permit the interaction between lincRNAs and AR. In the future we will validate these results by experimentally introducing point mutations into these motifs within the sequences of lincRNAs in order to understand their importance for lincRNA-AR binding.

Supported by FAPESP and Fundação Butantan.