Sequence analysis

**omiXcore: a web server for prediction of protein interactions with large RNA**

Alexandros Armaos,1,2 Davide Cirillo1,2 and Gian Gaetano Tartaglia1,2,3,*

1Bioinformatics and Genomics, Gene Function and Evolution, Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, and 2Bioinformatics and Genomics, Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain and 3Bioinformatics and Genomics, Institucio Catalana de Recerca i Estudis Avancats (ICREA), 08010 Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on December 30, 2016; revised on May 23, 2017; editorial decision on May 31, 2017; accepted on May 31, 2017

**Abstract**

**Summary:** Here we introduce omiXcore, a server for calculations of protein binding to large RNAs (> 500 nucleotides). Our webserver allows (i) use of both protein and RNA sequences without size restriction, (ii) pre-compiled library for exploration of human long intergenic RNAs interactions and (iii) prediction of binding sites.

**Results:** omiXcore was trained and tested on enhanced UV Cross-Linking and ImmunoPrecipitation data. The method discriminates interacting and non-interacting protein-RNA pairs and identifies RNA binding sites with Areas under the ROC curve > 0.80, which suggests that the tool is particularly useful to prioritize candidates for further experimental validation.

**Availability and implementation:** omiXcore is freely accessed on the web at http://service.tartaglia@lab.com/grant_submission/omixcore.

**Contact:** gian.tartaglia@crg.es

**Supplementary information:** Supplementary data are available at Bioinformatics online.

**1 Introduction**

RNA-binding proteins (RBPs) amount to a large number of heterogeneous molecules encompassing a vast array of biological functions and binding modalities (Marchese et al., 2016). The identification of RNA targets is important to characterize RBPs roles in physiological (Tartaglia, 2016) and pathological (Bolognesi et al., 2016) conditions. Considerable attention has been given to long non-coding RNAs that are implicated in important cell functions (Guttman and Rinn, 2012) but are difficult to characterize because of their tissue-dependent expression (Chen et al., 2016). Indeed, RNA interactions with RBPs require laborious experimental procedures such as chromatin isolation by RNA purification to detect protein networks bound to the RNA of interest (Chu et al., 2015). The development of enhanced UV Cross-Linking and ImmunoPrecipitation (eCLIP) has recently provided a wealth of information on RBPs-binding sites at the transcriptomic level (Van Nostrand et al., 2016). The large and homogeneous amount of data provided by eCLIP experiments represents an ideal dataset to train methods for prediction of protein interactions with long non-coding RNAs. Indeed, despite considerable efforts in RNA crystallography (Zhang and Ferré-D’Amare, 2014), the paucity of structural information leads to an urgency in the implementation of high-throughput approaches for identification of protein-RNA interactions. Using the catRAPID approach (Bellucci et al., 2011), we developed the uniform fragmentation procedure to predict interaction propensities between protein and RNA fragments (Cirillo et al., 2017). Here, we introduce omiXcore to perform predictions of long RNAs (500 nt and larger). Calibrated on eCLIP data, omiXcore allows fast and quantitative prediction of RBP interactions with human long intergenic RNAs (lincRNAs), facilitating experimental design and analysis.

**2 Workflow and implementation**

The omiXcore server allows calculation of the interaction propensities of a protein sequence against i) human lincRNAs (14717
entries available in http://www.ensembl.org/ or ii) a custom list of transcripts (maximum of 30 K characters). Once the user submits a protein of interest, the catRAPID signature algorithm (Livi et al., 2015) estimates the RNA-binding ability. If the protein is predicted to interact with RNA, its partners are calculated and the binding sites visualized.

- To train the algorithm, we used the eCLIP interactomes of 96 RBPs (56 studied in HepG2 and 78 in K562; downloaded from https://www.encodeproject.org in July 2016). We mapped targets of RBPs to their canonical transcript isoforms. For each RNA, we measured the overall affinity \( A^{\exp} \) defined as the number of reads (average of two replicas) divided by isoforms abundance (Trappnell et al., 2012).

- For each RBP, we ranked the transcripts by \( A^{\exp} \) and computed the local affinities \( A^{\exp} \) at each RNA site. To build the negative set, we compiled a list of transcripts that do not interact with the RBP of interest (i.e., they are not reported in the two eCLIP replicas) but bind to at least one of the other RBPs. In total we used 12,234 positive and 12,717 negative interactions (balanced set with 100 RNAs per RBPs).

- For each protein-RNA pair, we used the uniform fragmentation procedure to calculate interaction propensities between protein and RNA fragments (Cirillo et al., 2017). The uniform fragmentation approach is based on the division of protein and RNA sequences into overlapping segments [100 fragments for each molecule] (Cirillo et al., 2013). This analysis is particularly useful to identify protein and RNA regions involved in the binding.

- We computed mean \( \mu \) and SD \( \sigma \) of the interaction propensities for each RNA fragment \( i \) and the protein fragments, which we combined in the position-dependent vector \( \mathbf{F}_i = (\mu, \sigma) \).

- To predict the binding sites \( x_i^{\exp} \) of a specific RNA fragment \( i \), we integrated the interaction propensities \( F_i \) using the formula \( h_k = \tan(b(\omega_k F_i)) \) and calculating \( x_i^{\pred} = \tan(b(\Omega h_k)) \). Similarly, \( A^{\exp} \) is computed using \( h_k = \tan(b(\omega_k F_i)) \) and \( A^{\pred} = \tan(b(\Omega h_k)) \). Both \( x^{\pred} \) and \( A^{\pred} \) are defined in the range [0,1] and fitted to the experimental \( x^{\exp} \) and \( A^{\exp} \) optimizing the internal weights \( \omega_k \) and \( \Omega \) (neural network architecture with \( i = 100 \) and \( k = 50 \); total of 1.2 \( \times 10^6 \) binding regions used).

3 Performances

omiXcore builds on top of catRAPID algorithms that have been previously validated on a large number of interactions (Agostini et al., 2013; Cirillo et al., 2017; Livi et al., 2015): to evaluate omiXcore performances, we employed a leave-one-out procedure on the 96 individual subsets, each one corresponding to one RBP with its positive and negative interactors. Performances on RBP partners (Area under the ROC curve \( AUC = 0.83; \) Sensitivity = 0.75; Specificity = 0.78; Matthews correlation coefficient of 0.55, Fig. 1A) and RNA binding sites (\( AUC = 0.78; \) Sensitivity = 0.70; Specificity = 0.90; Fig. 1B) were assessed using a binary classification of interacting versus non-interacting pairs (at \( A^{\exp} \) and \( A^{\pred} \) cut-offs at 0.25). Cut-off points for \( A^{\pred} \) and \( A^{\pred} \) (0.5 and 0.1, respectively) were set maximizing the distance of the ROC curve from diagonal line (Fig. 1A and B). The 0.65 correlation (Spearman’s Rho) between \( x^{\exp} \) and \( x^{\pred} \) allows to quantify binding sites in the continuum range (Fig. 1B and C), which is useful to detect low-affinity interactions (Jankowsky and Harris, 2015). On the testing set, omiXcore shows higher AUCs (in the range of 0.93–0.99) than binary classifiers such as RPseq [RPseq-RF:0.50–0.60; RPseq-SVM:0.46–0.66] (Muppirala et al., 2011) and Global Score [0.55–0.88; see also Supplementary Material for other performances] (Cirillo et al., 2017).

4 Conclusions

In this work, we introduced the omiXcore tool for predicting RBP interactions with large RNAs. The algorithm allows detection of RNA binding sites by evaluating local physicochemical properties of polypeptide and nucleotide sequences (Bellucci et al., 2011). omiXcore was calibrated on eCLIP data (Van Nostrand et al., 2016) and is useful to prioritize coding and non-coding RNA targets for further experimental validation. We optimized the webserver to perform fast calculations of lincRNAs, for which we provide a pre-compiled library. Indeed, lincRNAs are poorly abundant and regulated in a precise spatiotemporal manner, which makes their characterization particularly difficult in the wet lab.

Acknowledgement

We would like to thank Fernando Cid for stimulating discussions.

Funding

We acknowledge support of the Spanish Ministry of Economy and Competitiveness, “Centro de Excelencia Severo Ochoa 2013-2017” and the CERCA Programme / Generalitat de Catalunya. This work was supported by the European Union Seventh Framework Programme [FP7/2007-13], European Research Council RIBOMYLOME_309545 (Gian Gaetano
Tartaglia) and Spanish Ministry of Economy and Competitiveness BFU2014-5505-P (Gian Gaetano Tartaglia).

Conflict of Interest: none declared.

References


