Functional microRNA targets in protein coding sequences

Martin Reczko, Manolis Maragkakis, Panagiotis Alexiou,
Ivo Grosse, Artemis G. Hatzigeorgiou

Merve Çakır

27.04.2012

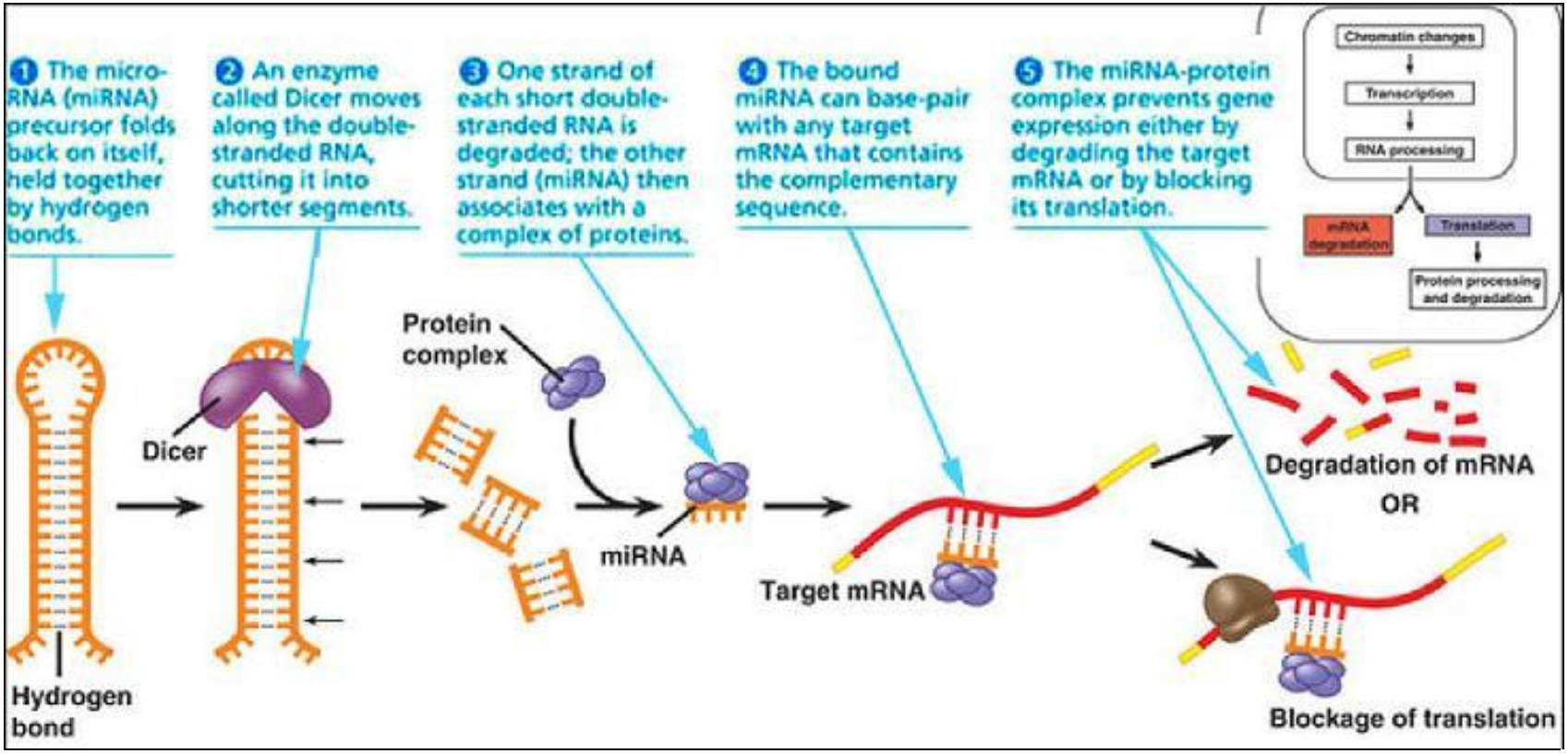
microRNA

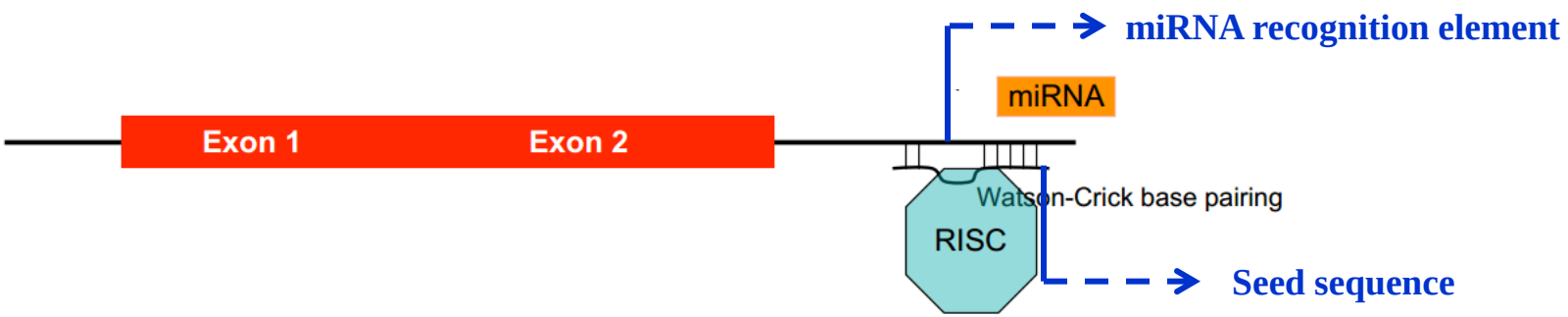
* microRNAs are small – 18-24 nucleotide long – noncoding RNAs whose function is to regulate expression of genes at mRNA level. By targeting complementary sequences on target mRNAs, microRNAs result in translational repression or degradation of target mRNA molecule.

* microRNAs take part in many biological processes ranging from development to cell proliferation and are involved in numerous diseases, including cancer.

* After its transcription, a microRNA molecule goes through some processing steps both in nucleus and cytoplasm and at the end, it is integrated into a multiprotein complex called RNA-induced silencing complex (RISC). microRNA molecules guide RISC to specific microRNA recognition elements based on binding specificity, after that RISC can take care of silencing the particular mRNA molecule.

miRNA Mechanism of Action





Background

Although most of the miRNA recognition elements have been found in 3'-UTRs of protein coding genes, recent studies imply that MREs can be located not just in 3'-UTRs but also within protein coding sequences of target genes.

* Forman *et al.* have shown the presence of four let-7 miRNA target sites within the CDS of the miRNA-processing enzyme Dicer, which can result in a mechanism for a miRNA/Dicer autoregulatory feedback loop.

* Wang *et al.* have shown that miR-107 tends to target sites within coding sequences but not in 3'-UTRs.

* A novel technique called PAR-CLiP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) was applied by Hafner *et al*. to analyze miRNA recognition elements in mRNA fragments. With this approach, they have identified that miRNAs tend to bind in approximately equal proportions on the 3'-UTR as well as on the protein coding sequences of target mRNAs.

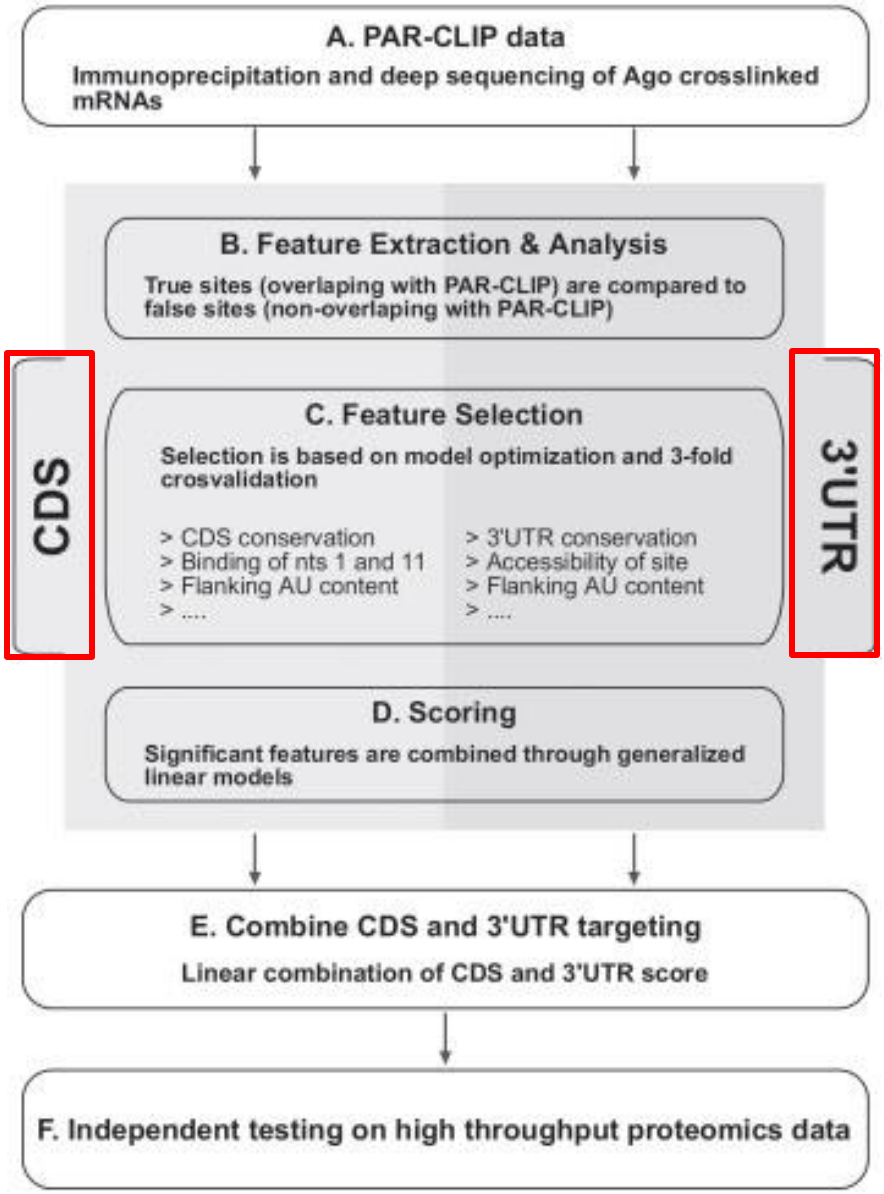
* After analyzing previously published high-throughput studies regarding miRNA targets, Fang and Rajewsky find that genes containing target sites both in the CDS and the 3'-UTR exhibit significantly stronger regulation than genes targeted in the 3'-UTR only and that this effect is stronger for conserved CDS sites with longer binding sites.

Motivation

All these studies show that CDS has as much potential as 3'-UTRs to contain a target sequence for miRNA binding and therefore, limiting the search for miRNA recognition elements only to 3'-UTR regions of mRNA sequences will not result in a complete identification of MREs.

Reczko *et al.* designed an algorithm for the prediction of miRNA targets in both 3'-UTRs and CDSs.

Flowchart of the Algorithm

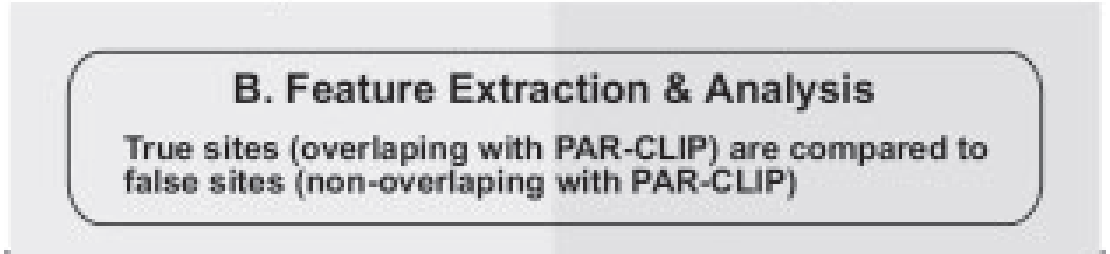


Feature Extraction

-A dynamic programming algorithm is used to identify the optimal alignment between the miRNA extended seed sequence and every 9 nucleotide window on the 3'-UTR or CDS.

- To identify the miRNA involved in each putative MRE position identified by Hafner *et al.*, sequences of all identified genomic locations of the PAR-CLIP data are aligned against the miRNA sequence of the top 100 expressed miRNAs. ➔ **true set**
All other aligned locations that do not overlap with the PAR-CLIP data ➔ **false set**

- 64 different binding categories are defined based on alignment procedures. These categories are then compared through a logistic regression between the binding categories and the presence or absence of the corresponding MRE in the true or false set of the PAR-CLIP data, in order to obtain "binding category weight" feature.

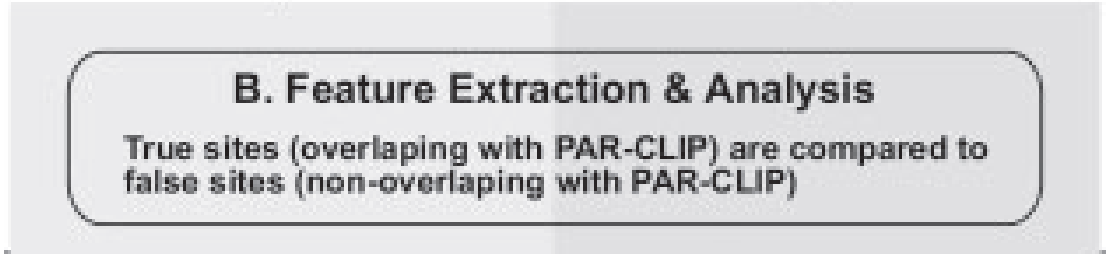


Feature Extraction

- A CDS conservation score is calculated for each MRE in CDSs based on the following reasoning: functional MREs in CDSs are expected to preferentially conserve those nucleotides that would have no effect on the amino acid outcome, but would interfere with miRNA targeting.

- Similarly, 3'-UTR conservation score is calculated for each MRE in 3'-UTRs, but this time evolutionary conservation of a MRE based on 16 species is used for scoring.

- Some other features such as MRE accessibility, flanking AU content, distance to closest 3'-UTR end, adjacent MRE distance and free energy of binding are also used.



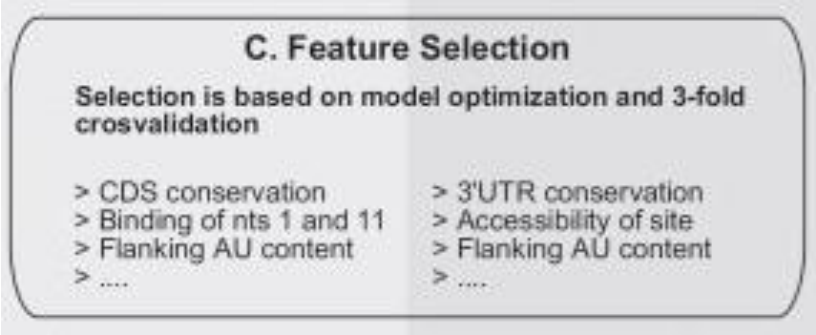
Feature Selection

Reczko *et al.* wanted to determine an optimal feature set using cross-validation, therefore the PAR-CLIP dataset is split into three disjoint subsets, stratified for positive and negative sites.

Logistic regression is then performed using all the previously described features on each subset and a feature selection procedure is used to determine the optimal set of features.

For this initial set of features, the capability of each single feature to separate the complete PAR-CLIP data into sites with reads and sites without reads is tested using the Wilcoxon's test and only features with significant separation are chosen.

* Feature selection is performed independently for sites in CDSs and sites in 3'-UTRs.

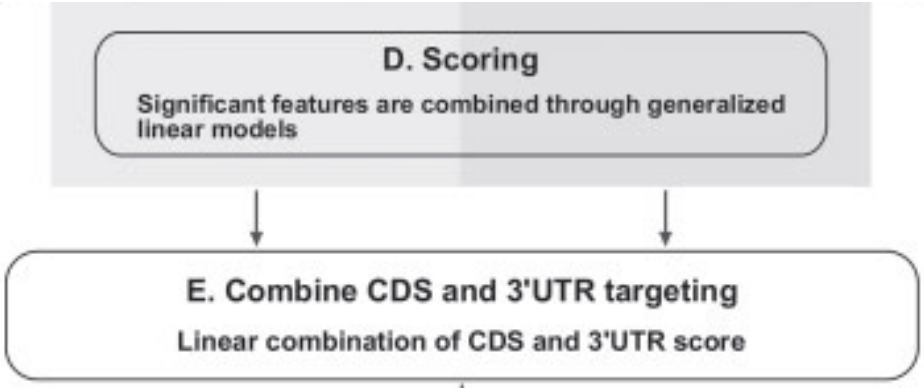


Training and Scoring

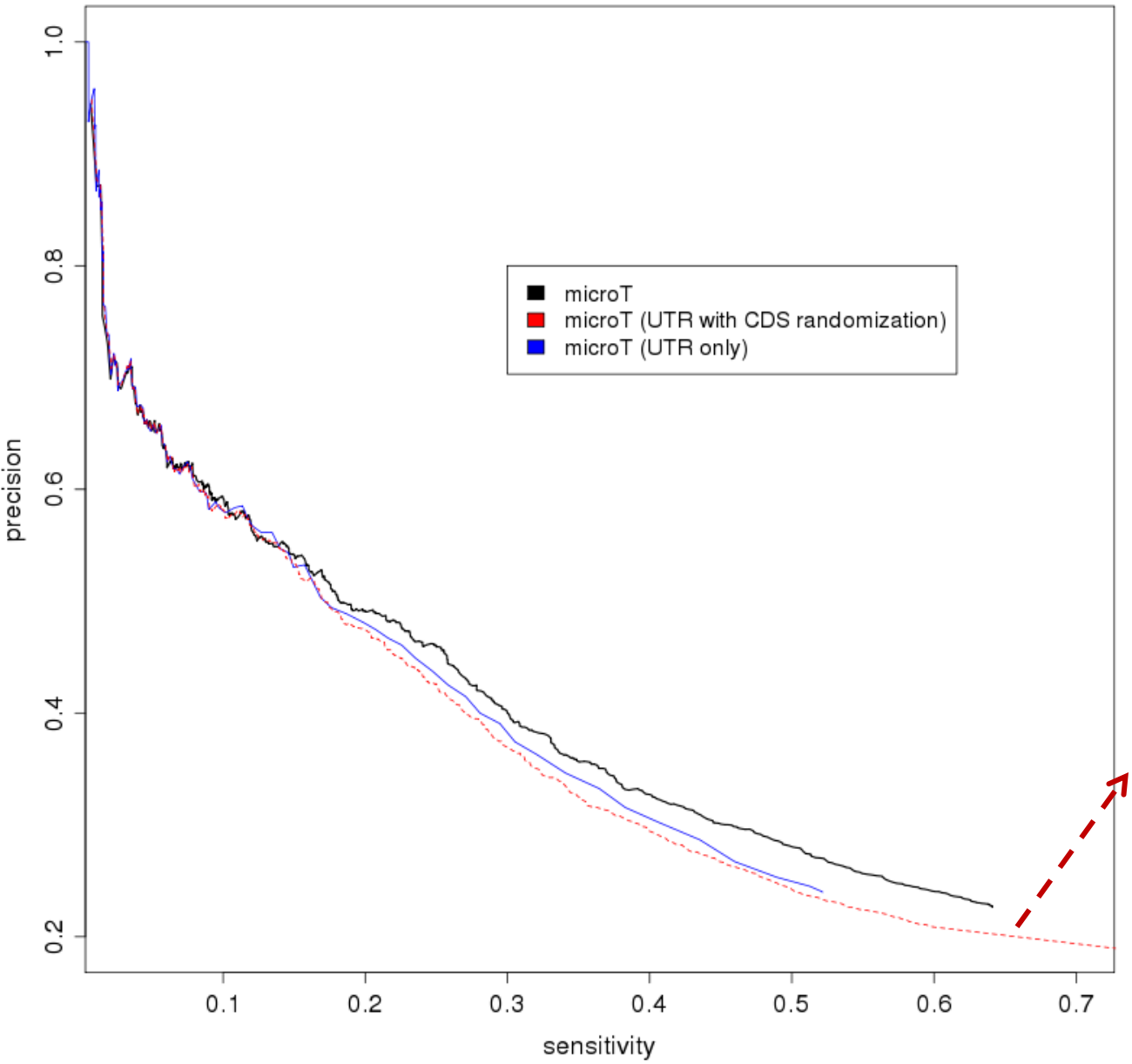
- Optimal set of features are then used in different machine learning approaches like support vector machines, neural networks, random forests and generalized linear models. MRE scores are calculated with each different method and the best performance, quantified by cross-validation, is obtained using generalized linear models.

Like feature selection, each gene region (CDS or 3'-UTR) is represented by a separate model. Scores for all MREs found in a region are then summed to obtain a region score.

- To combine CDS and 3'-UTR region scores, another generalized linear model is trained. Results of 13 different microarray experiments are used to generate true and false examples of the training set, instead of PAR-CLiP data.



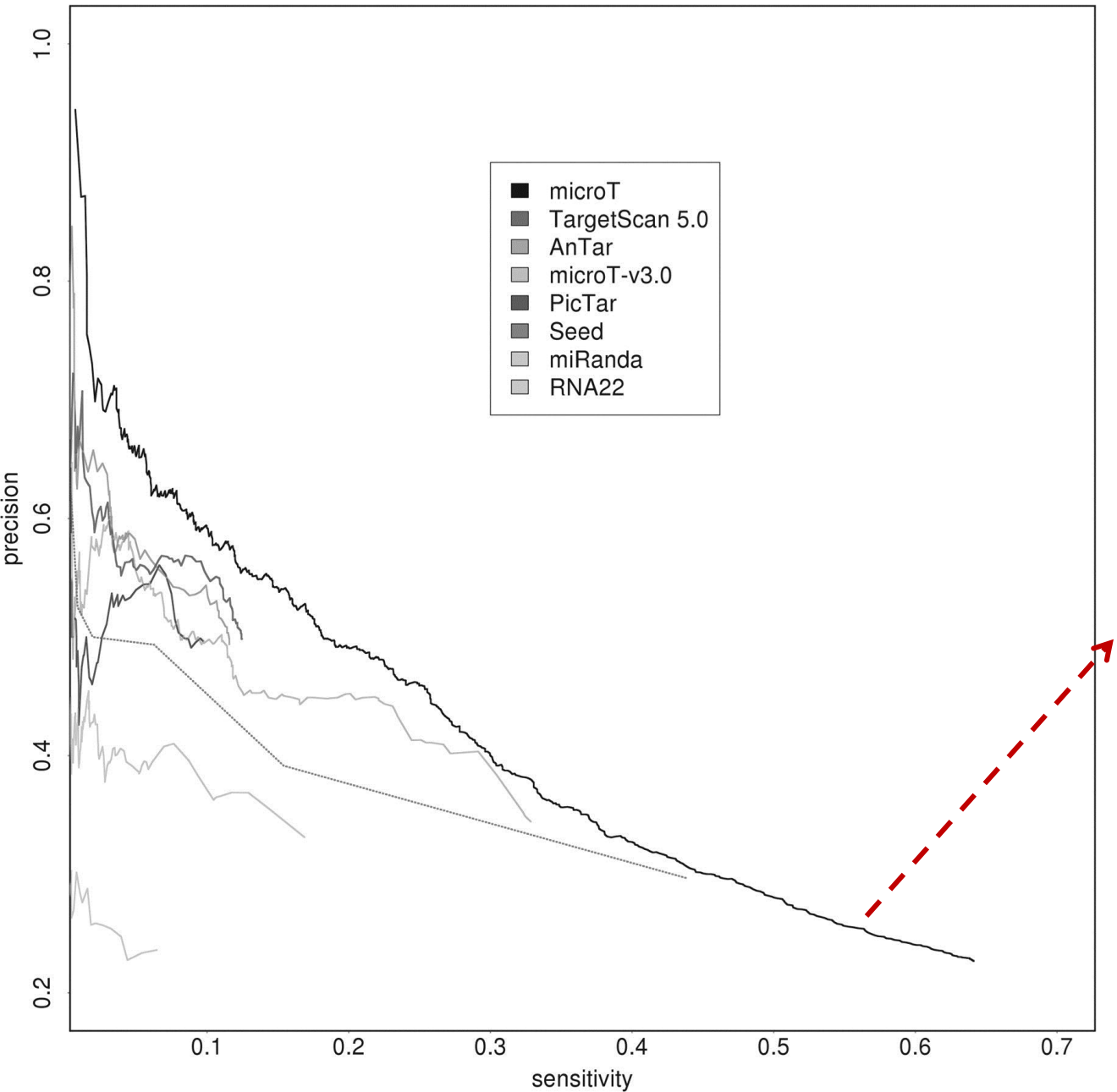
3’-UTR only vs 3’-UTR + CDS



* The combined model increases the sensitivity from 52% to 65% in comparison to the 3’-UTR only model, keeping the specificity at the same level of 32%.
➔ 293 additional correctly predicted targets

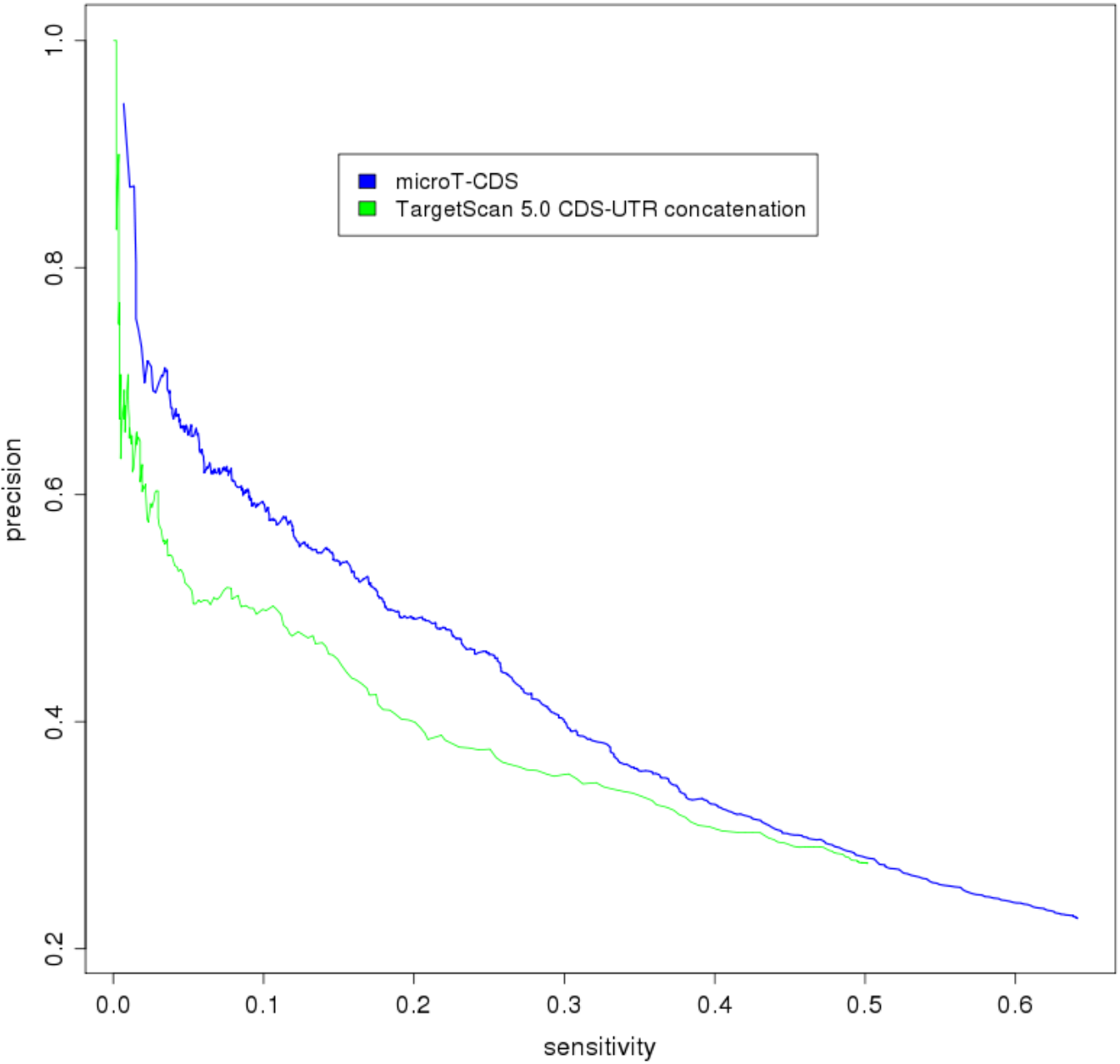
* The performance of randomized predictor is significantly lower than the combined model, demonstrating a significant and synergistic contribution of targeting in the CDS.

Comparison with Other Programs



microT-CDS exhibits the highest sensitivity at any level of specificity in comparison with the other programs.

Comparison with TargetScan 5.0

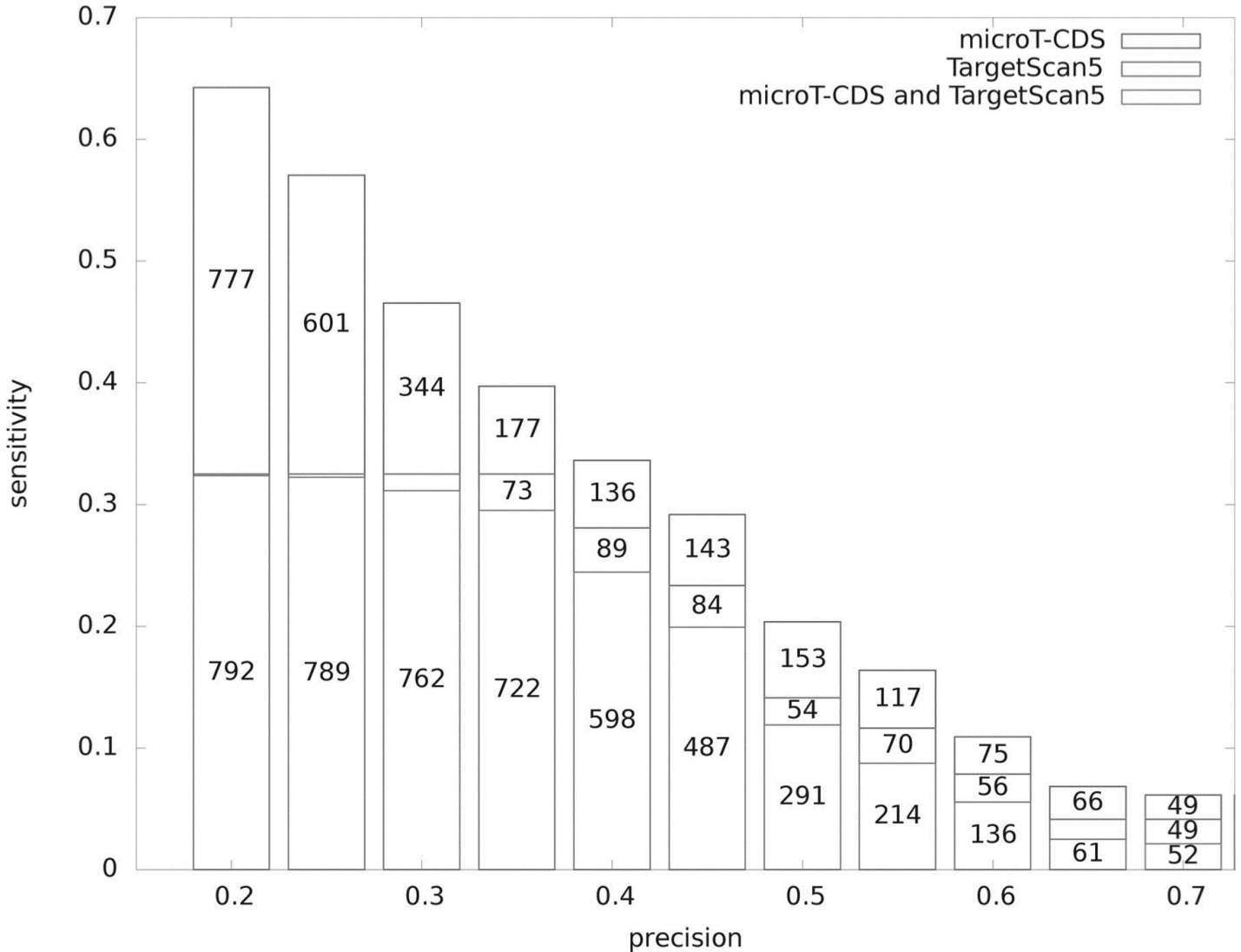


The validity of using a specific prediction model for the additional CDS sites is verified in a comparison with predictions of TargetScan 5.0 as this program can also use sites in CDS region.

* TargetScan 5.0 ended up with predictions having %10 lower precision compared to microT-CDS.

Comparison with TargetScan 5.0

- Depending on the precision level, the overlap between the targets predicted by microT-CDS and TargetScan 5.0 is found to be ranging from 50 to 70%.
- At lower precision levels the number of correct predictions is almost doubled using microT-CDS.



Additional Tests

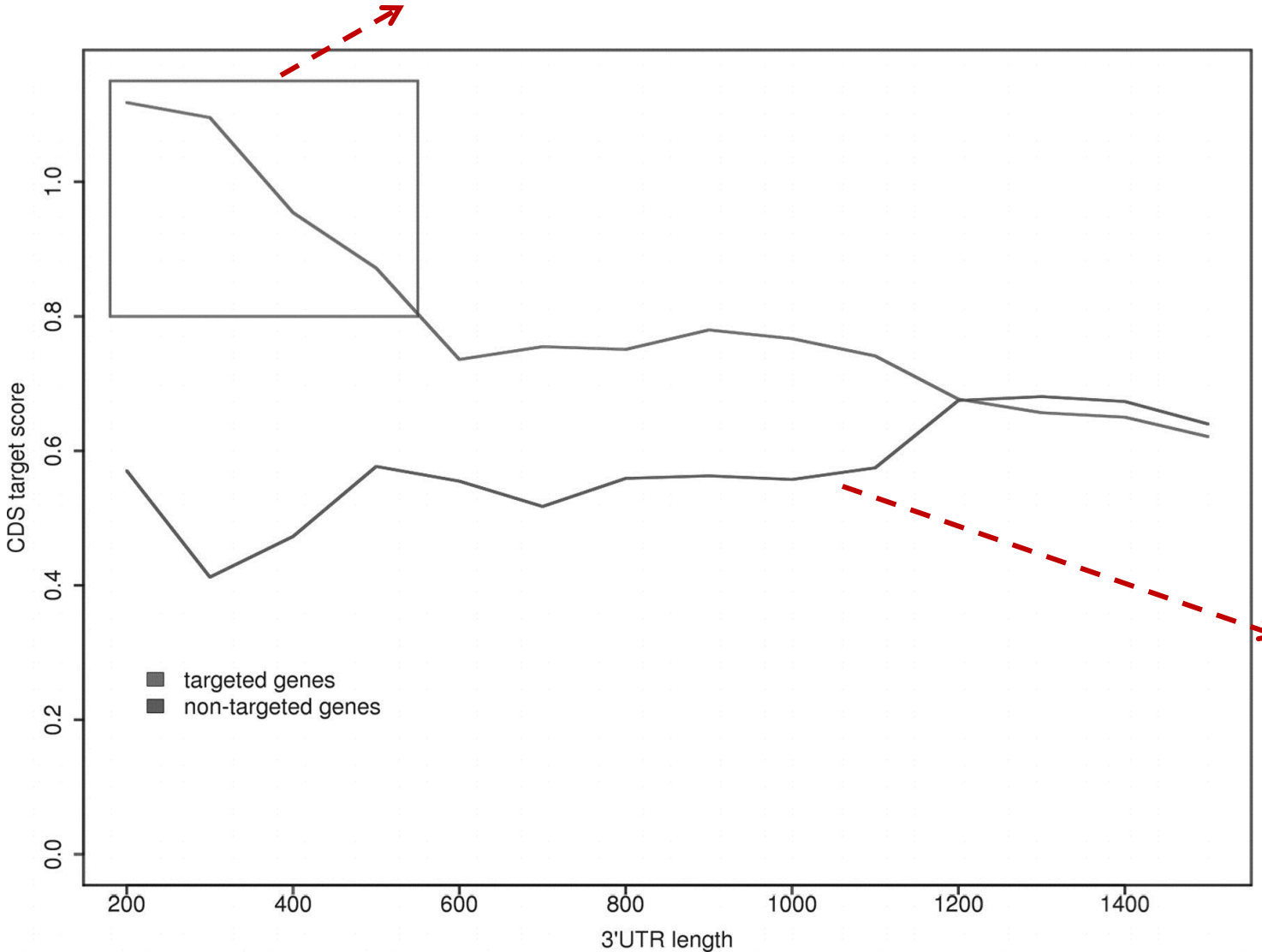
* The performance of microT-CDS in the detection of CDS target sites is also evaluated on the HITS-CLIP (high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation) dataset of Chi *et al.* microT-CDS is capable of predicting the location of 286 of 1210 target sites correctly.

To estimate if this prediction ratio could also be achieved by chance, the locations of the predicted sites is randomized 100 times. The randomized model is able to locate only 10.3 out of the 1210 real binding sites, leading to an estimated ratio of true over randomly predicted sites greater than 27.

* microT-CDS algorithm is also tested on five individual cases of experimentally verified MREs found in CDS regions and it was successful in recalling three of them, which is in agreement with the estimated sensitivity of the algorithm.

Effect of 3’-UTR Length

This region indicates all 3’-UTR lengths with significantly higher CDS sores, indicating likely targeting in the CDS.



* Genes having 3’-UTRs that are shorter than 500 nucleotides have a significantly higher CDS target score.

Such preference could not be observed for the group of genes that are measured as not targeted by miRNAs.

Conclusion

* Relationship between CDS targeting and 3'-UTR length suggests that evolutionary pressure might enforce the presence of additional sites on the CDS in cases where there is restricted space on the 3'-UTR.

* A feature analysis for MREs in 3'-UTR regions reveals a number of novel significant findings, such as the requirement for increased accessibility in the mRNA secondary structure at the start of an MRE.

* The analysis reveals also that functional MREs in the CDS preferentially require a stronger binding than MREs in the 3'-UTR. MREs in coding regions require a perfect binding along the miRNA seed region and mismatches disrupt their functionality.

The results of microT-CDS are available through the DIANA web server at www.microrna.gr/microT-CDS.

Thank You

Questions?