Improving microRNA target prediction by performance-based algorithm combination

Ignacio Sanchez Caballero[1], Ander Muniategui[2], Ruben Nogales-Cadenas[1], Carlos O. Sánchez-Sorzano[1], Angel Rubio[2] and Alberto Pascual-Montano[1]

[1]Functional Bioinformatics Group. National Center for Biotechnology-CSIC. Madrid, Spain
[2] CEIT and TECNUN, University of Navarra, San Sebastian, Spain.
E-mail: pascual@cnb.csic.es

CNB Centro Nacional de Biotecnología

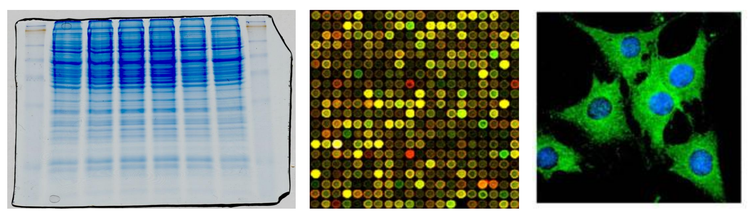
http://bioinfo.cnb.csic.es

Summary

The use of bioinformatics tools has become a major accelerator in our understanding of microRNAs function. Many algorithms have been created to predict where microRNAs are encoded, as well as what genes they regulate. Unfortunately, due to the popularity of the field, it is not always clear which of the available computational methods is best suited for determining which transcripts targets are regulated by which microRNAs. We propose a straightforward way to combine the tens of currently available prediction algorithms, and assign them a credibility measure based on their previous performance to simplify the task of experimental validation. Using some additional assumptions, we have created a new database, which provides a confidence score for each predicted interaction. This score is computed taking into account the number of databases where the interaction appears, the quality of these databases in terms of their predictive accuracy, and the ranking that each database assigns to its predictions. Using cross-validation, we show that this database outperforms in terms of quantity (number of interactions) and quality (ability to predict experimentally validated interactions) any of the previous ones. No algorithm makes perfect predictions under every condition. Because of the multi-faceted nature of miRNA targeting, and the lack of consensus among existing predictions, it makes sense to combine them in a way that maximizes the number of validated predicted results. There have been previous attempts to combine the predictions of several algorithms by first taking their union or intersection as a way to improve coverage or accuracy respectively, balancing out their sensitivity and specificity, and then choosing the most likely candidates by consensus. Most of these algorithms give the user the ability to choose which combination of databases should be used. The problem with this approach is that in a significant proportion of cases we do not have the necessary information about each database's performance to make an informed decision. Our approach presents an alternative solution that assigns confidence scores to each database's predictions. This solves the problem introduced by choosing candidates by consensus; mainly, that several low-confidence predictions for the same interaction can erroneously appear as more credible than a single high-confidence prediction.

Experimentally validated miRNA-mRNA interactions:

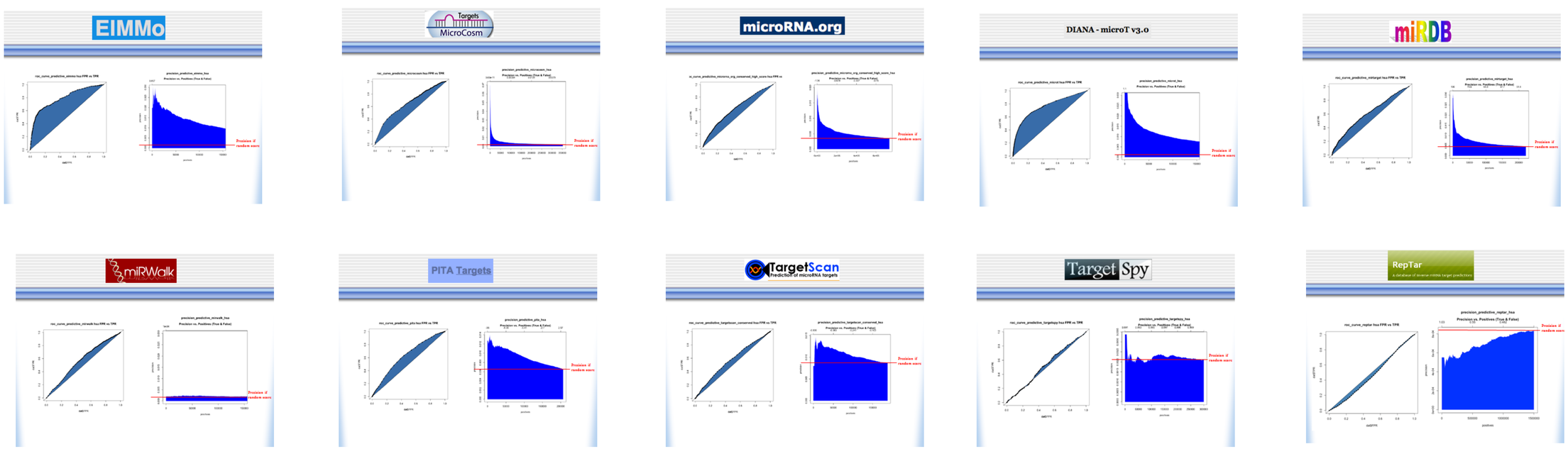
miRWalk · miRecords · miRTarBase · TarBase v.5c



In-silico predicted miRNA-mRNA interactions:

TargetSpy · PITA Targets · miRWalk · microRNA.org · Targets MicroCosm · EIMMo · TargetScan Prediction of microRNA targets · DIANA - microT v3.0 · miRDB · RepTar A database of inverse miRNA target predictions

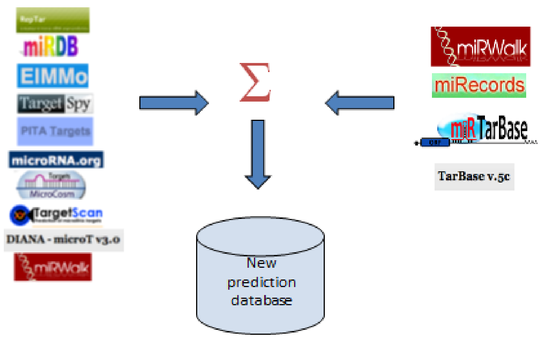
Performances of predictions taking validated predictions as gold standard (Homo Sapiens):



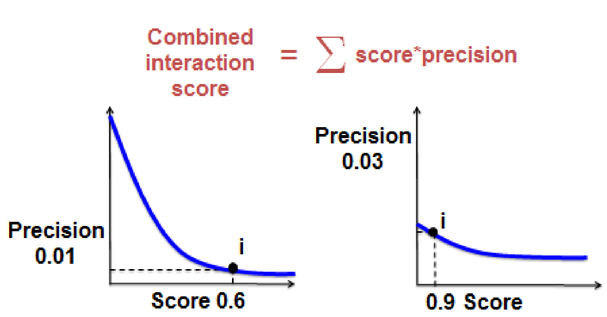
Integration of prediction algorithms:

Question: which one to use? **Answer:** Combine them!

Integration of predicting algorithms:



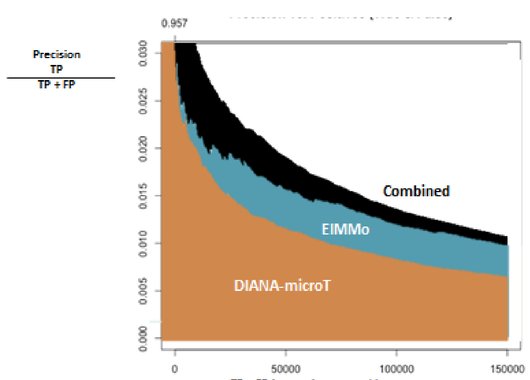
Idea: Interactions predicted by several algorithms are assigned a combined score



Integration of predicting algorithms:

- An initial attempt at combining all predictive databases might be to do it by score; that is, **sort** them by score, and in those cases where several databases predict the same interaction, take the **average score**. This is less than ideal, since each database defines a "good" score differently; for some, it might be anything above 0.6, for others nothing less than 0.9.
- Another attempt could be taking the **union**, or the **intersection** of all these algorithms. While unions of computational approaches may achieve a higher level of sensitivity than the individual approaches, this gain comes at the cost of a reduction in specificity. In the same vein, while intersections of computational approaches may achieve a higher level of specificity, they also generally achieve a much reduced sensitivity.

Algorithms perform better when combined



Steps:

1. First, normalize scores for each database (put them in the same context)
2. Sort the predictions by normalized score
3. Measure the credibility or precision of each database for each predicion: measures as the accumulated precision.
4. New score: Combine databases by score x precision for this prediction: different databses would support an interaction according to their "quality".

Conclusions

- Algorithm evaluation requires high-quality experimental datasets
- Databases are **divided by score regions** and evaluated on the number of validated interactions
- Algorithms perform much better when properly combined
- Simple union or interception won't solve the problem
- Predictions by sequence can also be combined with expression information to predict actual ineractions