Semi- supervised ensemble learning to boost miRNA target predictions

Gianvito Pio[1], Domenica D'Elia[2], Donato Malerba[1], Michelangelo Ceci[1] ✉
[1]Department of Computer Science - University of Bari, Bari, Italy
[2]CNR, Institute for Biomedical Technologies, Bari, Italy

Motivation and Objectives

The huge amount of data produced by the advent of Next Generation Sequencing (NGS) technologies is providing scientists with an unprecedented potential to investigate and shed light on remote secrets of genomes. In particular, many interesting insights are coming from the growing number of evidences about the regulatory function of the non-coding RNA (ncRNA) component of the genome. Indeed an increasing number of new ncRNAs have been recently discovered, most of them showing a primary role in the regulation of genome expression at different levels (Rossi Paschoal *et al.,* 2012). Some of these ncRNAs, although their existence is known since many years ago, have been only recently characterised for their functional role in important biological processes, in a wide variety of organisms and in human diseases. These findings represent one of the most important outcomes of the recent NGS revolution and it would not be so surprising if still unsuspected functions of this so called "dark matter" of the genome will be discovered in the near future.

Among functional classes of ncRNAs with a role in the regulation of gene expression, microRNAs (miRNAs) are those for which more functional data are available and on which the interest of scientists has been more focused over the last decades. The growing number of evidences of their key role in cancer and recent evidences about their presence in body fluids, such as serum and plasma, have further sparked the interest of the scientific community, emphasizing the possibility of using them as therapeutic targets and noninvasive biomarkers of disease and of therapy response.

We have developed a new tool based on biclustering techniques, i.e. HOCCLUS2 (Pio *et al.,* 2013) which is able to significantly correlate multiple miRNAs and their targets to detect potential miRNA:mRNA regulatory networks. However, experiments performed on predicted interactions led to observe that the noise (i.e., false positives) introduced by prediction algorithms can substantially affect the significance of the discovered modules. In order to overcome this issue, we have developed a probabilistic method which is able to build a more reliable dataset, combining data produced by several well-known miRNA target site prediction algorithms. This tool could greatly help in the interpretation of NGS miRNA profiles analysis with respect to their effects, by using genome-wide predictions of their targets.

Methods

The main goal of this work is to combine the prediction score of several prediction algorithms in a single stronger classifier, in order to improve the reliability of the obtained predictions. We propose a probabilistic approach to identify a probability function which, on the basis of the score returned by several prediction algorithms, estimates the probability of the presence of actual miRNA:mRNA interactions. In particular, it exploits information conveyed by datasets of validated interactions to learn such probability function. The identified function is then applied to large sets of predicted interactions. In this context, classical supervised learning algorithms cannot be directly applied, since: *i)* datasets of experimentally verified interactions provide only positive instances; *ii)* the number of labeled (positive) instances is imbalanced with respect to the number of unlabeled (unknown) instances. In order to overcome the first issue, the proposed approach works in the semi-supervised learning setting (Chapelle *et al.,* 2006) which exploits both information conveyed by (positively) labeled and unlabeled instances in the learning phase. Furthermore, the proposed method resorts to an ensemble learning solution in order to deal with the imbalancing.

The results obtained with the proposed approach, which "learns to combine" the output of several prediction algorithms, can be used to identify regulatory modules. This last task is performed by HOCCLUS2 which *i)* extracts possibly overlapping biclusters, to catch multiple roles of

both miRNAs and their target genes; *ii)* extracts hierarchically organized biclusters, to facilitate bicluster browsing and to distinguish between universe and pathway-specific miRNAs; *iii)* extracts highly cohesive biclusters, to consider only reliable interactions; *iv)* ranks biclusters according to the functional similarities, computed on the basis of Gene Ontology (GO)(Ashburner *et al.,* 2000), to facilitate bicluster analysis.

Results and Discussion

Experimental results obtained using human data from miRTarBase (Hsu *et al.,* 2011), as the set of labeled (positive) interactions, and mirDIP (Shirdel *et al.,* 2011), as the set of unlabeled (unknown) interactions, show a significant improvement in the quality of biclusters with respect to the baseline approach of averaging the scores obtained by selected prediction algorithms. In particular, the application of the proposed approach leads to identify biclusters containing miRNAs and mRNAs that are more functionally related each other, according to the GO classification. Moreover, a deep evaluation of the functional consistency of identified miRNA:mRNA modules, by investigating the current literature and data extracted from different web resources (e.g., DAVID, Reactome, GeneCards tool suite and STRING), shows performances of HOCCLUS2 which are comparable with those obtained when applied on experimentally validated interactions in miRTarBase.

Acknowledgements

This work is partial fulfillment of the research objective of "DM19410 - Laboratorio di Bioinformatica per la Biodiversità Molecolare" and "PON01 _ 02589 - MicroMap project "Caratterizzazione su larga scala del profilo metatrascrittomico e metagenomico di campioni animali in diverse condizioni fisiopatologiche".

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25-29. doi:10.1038/75556

Chapelle O, Schölkopf B, Zien A (2006) Semi-*Supervised Learning. Adaptive computation and machine learning.* MIT Press, Cambridge, Mass., USA.

Hsu SD, Lin FM, Wu WY, Liang C, Huang WC et al. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, **39**, 163-169. doi: 10.1093/nar/gkq1107

Pio G, Ceci M, D'Elia D, Loglisci C, Malerba D (2013) A novel biclustering algorithm for the discovery of meaningful biological correlations between miRNAs and mRNAs. *BMC Bioinformatics*, **14**(Suppl 7):S8. doi:10.1186/1471-2105-14-S7-S8

Rossi Paschoal AR, Maracaja-Countinho V, Setubal JC, Simoes ZLP, Verjovski-Almeida S, Durham AM (2012) Non-coding transcription characterization and annotation. A guide and web resource for non-coding RNA databases. *RNA Biology*, **9**, 274-282. http://dx.doi.org/10.4161/rna.19352

Shirdel EA, Xie W, Mak TW, Jurisica I (2011) NAViGaTing the Micronome - Using Multiple MicroRNA Prediction Databases to Identify Signalling Pathway-Associated MicroRNAs. *PLoS ONE*, **6**(2), e17429. doi:10.1371/journal.pone.0017429