BIOINFORMATICS LETTER TO THE EDITOR Vol. 30 no. 7 2014, pages 1047–1048
doi:10.1093/bioinformatics/btu100

Genome analysis

Advance Access publication March 10, 2014

Don’t use a cannon to kill the . . . miRNA mosquito

Nestoras Karathanasis[1,2], Ioannis Tsamardinos[3,4] and Panayiota Poirazi[2,*]

[1]Department of Biology, University of Crete, Heraklion, 71409, [2]Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Heraklion, 70013, [3]Department of Computer Science, University of Crete, Heraklion, 71409 and [4]Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), Heraklion, 70013, Greece

Associate Editor: Ivo Hofacker

Contact: poirazi@imbb.forth.gr

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

Received on September 6, 2013; revised on January 31, 2014; accepted on February 11, 2014

Supervised machine-learning approaches are frequently applied on biological data to learn a regression or classification model, whether used for prediction, classification or for gaining an understanding on the biological process that has generated the data (Larranaga *et al.,* 2006). Arguably however, it is sometimes the case that sophisticated and complicated methods are employed, published and advocated as advances without a comparison even against the simplest baseline methods. We consider a baseline method as the simplest method that an expert analyst can conceive within a few minutes of consideration of the problem and does not require any engineering or scientific ingenuity or novelty. A baseline can take the form of comparing against predicting by the mean of the outcome on a known dataset, without use of any special predicting variables, or comparing against random guessing. Because of this lack of comparison, the added-value of the sophisticated methods—if any—is not quantified; it remains unknown whether the extra effort for implementing or applying it is worth. A false perception about the difficulty of the problem may be created.

We now present an example of the above argument on the problem of identifying the position of miRNA mature molecules on their precursor RNA molecules, which typically have a hairpin-like secondary structure. In the cell, the miRNA precursor is first cut into a complex of two substring sequences (strands) with high complementarity called the ‘5′-strand’ and the ‘3′-strand’. The complex is called the miRNA:miRNA* duplex defined by its four corners denoted as k_{55}, k_{53}, k_{35} and k_{33} corresponding to the 5′-strand 5′-end, 5′-strand 3′-end, 3′-strand 5′-end and 3′-strand 3′-end positions, respectively (Fig. 1). The two strands are then separated and either one or both become a functional miRNA. The task is to predict the positions k_{55}, k_{53}, k_{35} and k_{33} given the sequence of a miRNA-hairpin molecule. Note that a hairpin corresponds to the sequence of the ‘Stem-loop’ structure as provided by miRBase, which does not necessarily represent the exact pre-miRNA sequence but consists of the latter extended with some flanking nucleotides. Solving the problem can suggest novel miRNAs within suspected miRNA-hairpin sequences, to guide miRNA discovery, as well as provide intuition regarding the mechanisms regulating the miRNA biogenesis.

The first tool that addressed this prediction task was ProMiR (Nam *et al.,* 2005) which did not use any baseline method. Subsequent studies used the previous methods as baselines without considering the simplest possible method. What is the baseline ‘straw man’ on this problem? Arguably, it is making a prediction based on the mean position of each corner k_{55}, k_{53}, k_{35} and k_{33} as estimated from a training set of known miRNA duplexes and their precursor sequences. We call this method the Simple Geometric Locator (SGL) obviously providing a constant predicted position on any hairpin independent of the input sequence. An important detail to address is to define the reference point for measuring the mean position since miRNA precursors have various lengths. We chose the terminal loop tip as the reference point as it does not depend on the length of the pre-miRNA-flanking regions included in the hairpin sequence (see Supplementary Material, file 1 for details).

The set of ‘cannons’ to compare against SGL form four of the state-of-the-art tools for the task, namely MatureBayes [we note that MatureBayes did compare against the SGL in Gkirtzou (2009). Unfortunately, the reference point used for the SGL was the beginning of the flanking regions, whose length is arbitrarily chosen before miRNA precursors are inserted in the MiRBase; hence, the performance of the SGL was found inferior in that work] (Gkirtzou *et al.,* 2010), MiRPara (Wu *et al.*, 2011), MaturePred (Xuan *et al.*, 2011) and the most recent MiRdup (Leclercq *et al.,* 2013) published in respectable venues such as *PLoS ONE*, *BMC Bioinformatics* and *Nucleic Acids Research*. These tools employ machine-learning algorithms such as the Naïve Bayes Classifier, Support Vector Machines and the random forest classifier. They also employ complex raw and constructed features that include the nucleotide sequence, the secondary structure, number of loops and bulges, matches or mismatches for each nucleotide and others.

In our comparison, ‘prediction error’ on the task for each corner (end) is measured as the End Absolute Error (EAE): the absolute error of the predicted minus the true position (in nucleotides) for a specific duplex end (see Supplementary Material, file 1 for an example). To measure ‘prediction accuracy’, we define as ‘correct’ a prediction with error less or equal to a number x, i.e. EAE $\leq x$. Then, the prediction accuracy for an error bound (tolerance) of at most x, denoted as Accu(x), is the percentage of correct predictions in the test set. For example, if a

*To whom correspondence should be addressed.

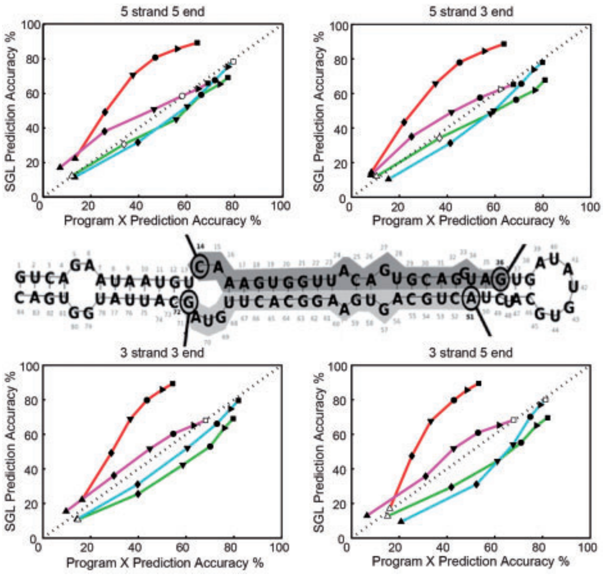

Fig. 1. Prediction performance per corner. Performance accuracies are estimated using the EAE for up to 5 nt. In each subplot, the y-axis shows the prediction accuracy of the Simple Geometric Locator (in %) and the x-axis shows the prediction accuracy of other methods (in %) for the same error tolerance. Lines comparing against the Simple Geometric Locator correspond to MatureBayes (magenta), MaturePred (red), MiRPara (cyan) and MiRdup (green). In addition, upward triangle, diamond, downward triangle, circle, right pointed triangle and square correspond to errors ≤ 0, 1, 2, 3, 4, 5 nt, respectively. Statistically significant results are indicated with filled symbols

model identifies the position of a given duplex end in 50% of duplexes within at most ±4 nt from their true position, it has accuracy at 4 nt of 50%: Accu(4) = 0.5. 'Statistical significance' of the results is assessed by assuming the null hypothesis that two methods have the same accuracy for a given error bound and applying the Fisher's exact test. To ensure fairness, in each comparison the SGL is trained (estimates the mean positions) with each method's training set, the one employed in the corresponding publication, after removing all miRNA hairpins with unknown duplexes. The performance accuracies of all tools are estimated on a common hold-out test set, as detailed in Supplementary Material, file 1. Since some tools do not provide a prediction on all hairpins, the estimation of accuracy is computed only on the hairpins for which a prediction is made; SGL of course, always provides a prediction.

The results are shown in Figure 1. First, the accuracies $\mathrm{Accu}_i(x)$ are computed for each tool i and duplex end for an EAE of $x = 0$, 1, 2, ..., 5nt. Subsequently, these accuracies are plotted against each other by connecting the points ($\mathrm{Accu}_i(0)$, $\mathrm{Accu}_{\mathrm{SGL}}(0)$), ($\mathrm{Accu}_i(1)$, $\mathrm{Accu}_{\mathrm{SGL}}(1)$), ..., ($\mathrm{Accu}_i(5)$, $\mathrm{Accu}_{\mathrm{SGL}}(5)$). For example, a point ($\mathrm{Accu}_i(1)$= 30%, $\mathrm{Accu}_{\mathrm{SGL}}$ (1)= 40%) implies that method i identified 30% of the duplexes in the test set within ±1 nt of their true position, while SGL identified 40% of duplexes within ±1 nt of their true position. Thus, if a line is on the diagonal, then the two methods achieve the same accuracy for the same error tolerance. If it is below the diagonal SGL achieves lower accuracy for the same error tolerance and if it is above the diagonal, then the SGL achieves higher accuracy for the same error tolerance than the method compared against.

As evident from the figure, SGL clearly and statistically significantly outperforms MaturePred (red) in predicting any of the four duplex corners for all error bounds and MatureBayes (magenta) for error bounds up to 3–4 nt. On the 5′-strand SGL and MiRdup (green) achieve similar accuracies for absolute error of at most 0 and 1 nt with MiRdup slightly improving for larger error bounds; MiRPara (cyan) on the 5′-strand is better by only 2–5%. In the 3′-strand, MiRdup and MiRPara exhibit an overall better performance than SGL. However, when focusing on the accuracy with zero tolerance Accu(0), i.e. the percentage of duplexes identified on their exact position corresponding to the first point of each line, only MiRPara shows significantly better results with the difference in performance ranging from 4–11% (see Fig. 1 and Supplementary Table S1).

CONCLUSION

Comparing against the simplest possible method as a baseline in data analysis is an important step of the analysis. Foregoing this step may result in unnecessary effort and energy spent in code developing, publishing and evaluations by future researchers, unnecessary use of computationally expensive methods and a false impression about their benefits and added value they provide in a given task. As an example we show that using the mean positions in predicting the four corners of a miRNA duplex complex outperforms some state-of-the-art methods and is on par with the rest when trying to predict the exact location of the duplex with zero tolerance.

Funding: European Union (European Social Fund—ESF); Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF)—Research Funding Program: Heracleitus II and by the BIOSYS research project, Action KRIPIS, project No MIS-448301 (2013SE01380036) that was funded by the General Secretariat for Research and Technology, Ministry of Education, Greece and the European Regional Development Fund (Sectoral Operational Programme: Competitiveness and Entrepreneurship, NSRF 2007-2013)/ European Commission.

Conflict of Interest: none declared.

REFERENCES

Gkirtzou,K. (2009) Mature MiRNA Identification via the use of naive Bayes classifier. Master Thesis, University of Crete, Heraklion, Crete, Greece.

Gkirtzou,K. *et al.* (2010) MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors. *PloS one*, **5**, e11843.

Larranaga,P. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinform.*, **7**, 86–112.

Leclercq,M. *et al.* (2013) Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucl. Acids Res.*, **41** (15), 7200–7211.

Nam,J.W. *et al.* (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.*, **33**, 3570–3581.

Wu,Y. *et al.* (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinform.*, **12**, 107.

Xuan,P. *et al.* (2011) MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. *PloS one*, **6**, e27422.