Comparison of Four Ab Initio MicroRNA Prediction Tools

Müşerref Duygu Saçar and Jens Allmer

Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey

Keywords: miRNA, Ab Initio, miRNA Prediction, Comparison.

Abstract: MicroRNAs are small RNA sequences of 18-24 nucleotides in length, which serve as templates to drive post transcriptional gene silencing. The canonical microRNA pathway starts with transcription from DNA and is followed by processing by the Microprocessor complex, yielding a hairpin structure. This is then exported into the cytosol where it is processed by Dicer and next incorporated into the RNA induced silencing complex. All of these biogenesis steps add to the overall specificity of microRNA production and effect. Unfortunately, experimental detection of microRNAs is cumbersome and therefore computational tools are necessary. Homology-based miRNA prediction tools are limited by fast miRNA evolution and by the fact that they are template driven. Ab initio miRNA prediction methods have been proposed but they have not been analyzed competitively so that their relative performance is largely unknown. Here we implement the features proposed in four microRNA ab initio studies and evaluate them on two data sets. Using the features described in Bentwich 2008 leads to the highest accuracy but still does not provide enough confidence into the results to warrant experimental validation of all predictions in a larger genome like the human genome.

1 INTRODUCTION

MicroRNAs (miRNAs) are a group of small noncoding RNAs, discovered in the early 90s by Ambrost and colleagues (Lee et al., 1993), which convey posttranscriptional regulation. In most cases miRNAs lead to down regulation of their target mRNAs but translational activation has also been observed (Orom et al., 2008). It has been estimated that 60% of all human genes are regulated by miRNAs (Friedman et al., 2009). Another estimate is that there are more than 1000 miRNAs in the human genome (Berezikov et al., 2005). MiRNAs can come from introns (Morlando et al., 2008), coding regions (Rodriguez et al., 2004), or intergenic miRNA gene clusters (Altuvia et al., 2005). It has been suggested that a miRNA may regulate hundreds of targets (Enright et al., 2003). MiRNAs can, therefore, form complex regulatory networks. Not surprisingly, miRNAs are implicated in diseases such as cardiovascular disease (Elton et al., 2011) and cancer (Suzuki and Miyazono, 2011). The biogenesis of miRNAs follows largely the canonical pathway. Initially, DNA is transcribed into RNA by either RNA polymerase II (Lee et al., 2004) or III (Borchert et al., 2006) and then the microprocessor complex (Han et al., 2006) cleaves hairpin structures from the transcript. These hairpins are exported into the cytosol by Exportin 5 (Lund et al., 2004; Zeng and Cullen, 2004; Okada et al., 2009) where they are cleaved by Dicer (Cifuentes et al., 2010) and then loaded onto the RNA induced silencing complex.

Despite the great effort that has been put into the elucidation of the microRNA pathway, not much is known which would facilitate computational modeling that is based on clear processing facts instead of data mining approaches. Two approaches are available for microRNA prediction, one based on homology and the other free of any references named ab initio.

While homology-based methods seem straightforward, they only retrieve results similar to already known microRNAs and rarely allow the detection of novel microRNAs (Bentwich et al., 2005). Furthermore, microRNA evolution progresses at a high rate (Liang and Li, 2009); (Lu et al., 2008), which limits the applicability of homology-based methods. Ab initio prediction methods (Ng and Mishra, 2007); (Lai et al., 2003); (Bentwich, 2008); (Ding et al., 2010); (Jiang et al., 2007); (Pfeffer et al., 2005); (Xue et al., 2005); (Yousef et al., 2006); (Grundhoff, 2011); (Burgt et al., 2012); (Cakir and Allmer, 2010); (Ritchie et al., 2012) try to extract parameters which describe hairpin structures, an element which...
is deemed important in the miRNA genesis process, and uses these features for machine learning to detect miRNAs. Although many of the \textit{ab initio} algorithms that have been proposed report their accuracy, they cannot be easily compared as they are run on different data sets. Furthermore, most of the algorithms cannot be obtained. Therefore, we implemented all features described in Ding et al., Jiang et al, Ng and Mishra, and Bentwich (Ng and Mishra, 2007); (Bentwich, 2008); (Ding et al., 2010); (Jiang et al., 2007) and compared them on the same data set to investigate relative algorithm performance. Ding et al. performed best on the data sets which we prepared but the maximum achieved accuracy of 0.996 would produce too many false positives so that experimental validation of all predictions would not be cost and time effective. Furthermore, this high accuracy seems to be an outlier and in all other cases that we tested Bentwich 2008 outperforms Ding et al. 2010 with a maximum accuracy of 0.986 that is closely reproduced among data sets. Therefore, we advise the use of the features described in Bentwich 2008 when attempting \textit{ab initio} hairpin prediction.

2 MATERIALS AND METHODS

The four tools which we wanted to compare are not available as software so we implemented all features that were proposed in the papers in Java™ and used our code to calculate the values from the negative and positive data sets.

2.1 Features

The features that are used to discriminate between a true miRNA hairpin and a pseudo one are different among studies, but sometimes the feature is the same and just the naming differs. In the following we list the features that were used in the studies compared here and add the acronym that we gave to the feature in parentheses. Features are also summarized in Table 1 for ease of reference.

2.1.1 Features Used in Ng and Mishra 2007

Sequence based features; 16 dinucleotide frequencies %NN (%AA, %AC, %AG, %AU, %CA, %CC, %CG, %CU, %GA, %GC, %GG, %GU, %UA, %UC, %UG, %UU) and 1 aggregate dinucleotide frequency %G+U (%G++%C).

Probability based features derived from dinucleotide shuffling (dns); adjusted base pairing propensity \(dp\) (dns\(_p(bpp)\), dns\(_p(bpp/hpl)\)), adjusted Minimum Free Energy of folding (MFE) \(dG\) (dns\(_p(hpmfe_rf)\), dns\(_p(hpmfe_{rf/hpl})\)), MFE index 1 MFEI1 (hpmfe\(_{rf/11}\)), adjusted base pair distance \(dD\) (dns\(_p(bpd)\), dns\(_p(bpd/hpl)\)), adjusted shannon entropy \(dQ\) (dns\(_p(Q)\), dns\(_p(Q/hpl)\))_.

2.1.2 Features Used in Jiang et al., 2007


2.1.3 Features Used in Bentwich, 2008

Structural features; hairpin length (hpl), loop length (hll), free energy per nucleotide (hpmfe\(_{rf/hpl}\)), matching base pairs (bpp) and maximal bulge size (mb).

Sequence based features; abundance of any dinucleotide, AA, AT, etc. (#AA, #AC, #AG, #AU, #CA, #CC, #CG, #CU, #GA, #GC, #GG, #GU, #UA, #UC, #UG, #UU), regular internal repeat (dr), inverted internal repeat (ir), free energy (hpmfe\(_{rf}\)) and GC content (%GC).

2.1.4 Features Used in Ding et al., 2010


Sequence based features; base pairing propensity \(dp\) (bpp), \(dp/nloops\) (bpp/nl), \(avgbpstem\) (bpp/sl), diversity (bpp), \(|A-U|/L\) (%AU), \(|G-C|/L\) (%GC), \(|A-U|/nloops\) (st(A-U)/ns), \(|G-C|/nloops\) (st(G-C)/ns).

Thermodynamics based features; ensemble free energy NEFE (efe), minimum free energy index 1-4 MFEI1 (hpmfe\(_{rf/11}\)), MFEI2 (hpmfe\(_{rf/ns}\)), MFEI3 (hpmfe\(_{rf/ns/hpl}\)), MFEI4 (hpmfe\(_{rf/hpl}\)), dG (dG), Diff (dme), ensemble frequency Freq (efq), melting temperature Tm (Tm), enthalpy divided by length dH/L (dH/hpl), entropy divided by length dS/L (dS/hpl), Tm/L (Tm/hpl), p-value_MFE (dns\(_p(hpmfe_{rf})\), p-value_EFE (dns\(_p(efe)\), z-core_MFE (dns\(_z(hpmfe_{rf})\), z-core_EFE (dns\(_z(efe)\)).
Table 1: Features that were proposed in the selected studies are presented in the first row of the respective study and the acronyms we chose for those features are presented in the following row.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Sequence-Based</th>
<th>Probability-Based</th>
<th>Structural</th>
<th>Thermodynamic-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ng and Mishra</td>
<td>16 dfs %NN and 1 aggregate df%G+C ratio</td>
<td>dP, dG, MFEI1, dD, dQ, MFEI2, dF, normalized variants of dP, dG, dQ, dD and dF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>%AA-%UU, %G+C</td>
<td>dns_p(bpp, bpp/hpl, hpmfe_rf, hpmfe_rf/hpl, bpp, bpp/hpl, Q, Q/hpl), dns_z(bpp, bpp/hpl, hpmfe_rf, hpmfe_rf/hpl, bpp, bpp/hpl, Q, Q/hpl), hpmfe_rf_I1, hpmfe_rf/hs, dc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jiang et al. 2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bentwich 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ding et al. 2010</td>
<td>base pairing propensity (dP), Avg_bp_stem, diversity,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hpl, hll, (hpmfe_rf/hpl), bpp, mbs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2 Data Sets

2.2.1 Positive Data Set

MirBase is the de facto standard repository for miRNAs (Griffiths-Jones et al., 2008). It contains about 1500 entries for human counting both guide and passenger strands. We downloaded all human miRNAs as positive data. From the entries we removed the ones that contain more than one hairpin when folded by RNAfold (Hofacker, 2003) or RNAshapes (Steffen et al., 2006). If no proper link to Ensemble could be established, the entries were removed as well. From the remaining, about 1000, miRNA examples we created five random subsets containing 500 positive examples each.

2.2.2 Negative Data Sets

Negative data sets are especially difficult to establish for miRNAs, experimentally or computationally (Ding et al., 2010); (Ritchie et al., 2012); (Wu et al., 2011); (Yousef et al., 2008). Since most machine learning approaches that have been proposed for ab initio miRNA prediction are built on two class classification we designed one data set which consists of random sequences of the same length as the selected miRNAs in the positive data set. We
consider this data set to be easy to solve. Another data set, the pseudo miRNA data set, was taken from the Ng and Mishra study and we consider it to be more difficult (Ng and Mishra, 2007).

2.3 Machine Learning

We created five combined data sets consisting of 60% training and 40% test data from the overall data set. These data sets were used to train and test SVM classification using default settings in Orange Canvas (http://orange.biolab.si/). For performance evaluation test on ‘test data’ was used. This approach was used since fivefold cross validation could not be used with multiple studies at the same time; but had to be repeated individually, thus leading to different data sets and therefore to a potentially unfair comparison.

3 RESULTS AND DISCUSSION

In our opinion, the two datasets that were used (see Data Sets) are of different difficulty with the random dataset being easier to solve than the pseudo miRNA data set. This can also be deduced from the best results reported in Tables 2 and 3. The results for the random miRNAs in Table 2 lead to higher accuracy than the data in Table 3 which is achieved with pseudo miRNAs. For both tables the best and the average accuracy are provided along with the standard deviation, calculated from fivefold cross validation. Using the features described in the study by Bentwich 2008 leads to the highest accuracy without any standard deviation. Judging from these results it seems trivial to discriminate between true and false hairpins when using properly selected features.

Table 2: Accuracy measurements for human miRNAs (positive dataset) and random miRNAs (negative dataset).

<table>
<thead>
<tr>
<th>Studies</th>
<th>Best</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ng and Mishra 2007</td>
<td>0.919</td>
<td>0.894</td>
<td>0.183</td>
</tr>
<tr>
<td>Bentwich 2008</td>
<td>1.000</td>
<td>1.000</td>
<td>0</td>
</tr>
<tr>
<td>Ding et al. 2010</td>
<td>1.000</td>
<td>0.676</td>
<td>0.217</td>
</tr>
<tr>
<td>Jiang et al. 2007</td>
<td>0.954</td>
<td>0.952</td>
<td>0.003</td>
</tr>
</tbody>
</table>

We did not expect such a perfect result as achieved by Bentwich 2008 features in Table 2 for the pseudo miRNAs and Table 3 displays no such success. For the pseudo miRNA dataset, the features used in Ding et al. 2010 achieve the highest accuracy although with a high standard deviation over the cross validation.

Table 3: Accuracy measurements for human miRNAs (positive dataset) and pseudo miRNAs (negative dataset).

<table>
<thead>
<tr>
<th>Studies</th>
<th>Best</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ng and Mishra 2007</td>
<td>0.930</td>
<td>0.895</td>
<td>0.060</td>
</tr>
<tr>
<td>Bentwich 2008</td>
<td>0.986</td>
<td>0.983</td>
<td>0.002</td>
</tr>
<tr>
<td>Ding et al. 2010</td>
<td>0.996</td>
<td>0.599</td>
<td>0.198</td>
</tr>
<tr>
<td>Jiang et al. 2007</td>
<td>0.910</td>
<td>0.877</td>
<td>0.018</td>
</tr>
</tbody>
</table>

The best accuracy of 0.996, achieved by the features described in the study by Ding et al. 2010, when used for a million putative hairpins in human would lead to 4000 false positive identifications. Unfortunately, the number of putative hairpins in human is large and the accuracy calculated here does not fully reflect the true accuracy.

This is due to the fact that it is not entirely known what differentiates a true from a false hairpin. A much higher false positive rate must therefore be expected for real data and thus the number of false positives may not allow costly experimental validation of all predicted miRNAs.

Figures 1 and 2 further support this point by showing that the accuracy strongly depends on the data set used for training and testing. This can be deduced from the variation among the accuracies for the 5 data sets used in the fivefold cross validation.

For the random miRNAs, the variation among datasets is large for most studies. However, Bentwich 2008 always achieves perfect separation and Jiang et al. 2007 achieves a low variation and an overall good result (Figure 1).

With the pseudo miRNAs the variation is even more important to be analysed. Although Ding et al. 2010 achieves the highest accuracy in one case, it fails in all other cases which shows a strong dependence on the training and test data set and a poor generalization for the features from that study (Figure 2). Bentwich 2008 does not have such generalization problems and outperforms all other
studies on the remaining four data sets.

![Figure 2: Accuracy measurements for human miRNAs (positive dataset) and pseudo miRNAs (negative dataset). All cross validation results are shown individually.](image)

4 CONCLUSIONS

Although many algorithms have been proposed for ab initio miRNA gene prediction, they have not been compared for their relative performance. We compared four of twelve available ab initio algorithms in this study and found that Bentwich 2008 achieves the highest accuracy on the random data set and the second best accuracy on the pseudo hairpin data set but with a very low variation. Unfortunately, the achieved accuracy of 0.986 would lead to many false positives which would turn any attempt at experimental validation of all predicted miRNAs into a futile endeavor.

In the future, we plan to expand this assessment of available algorithms to all currently available ones. We believe it is necessary to establish the accuracy of existing algorithms not independently but transparently and comparable. To the best of our knowledge, this is the first independent assessment of multiple ab initio miRNA prediction methods.

As negative data sets are hard to come by, we will try to establish another set of negative data and further try one-class classification with the proposed parameters in follow-up studies.

Currently, we advise the use of the features used in the Bentwich 2008 study when trying ab initio prediction of hairpins.

ACKNOWLEDGEMENTS

This study was in part supported by an award received from the Turkish Academy of Sciences (http://www.tuba.gov.tr) for outstanding young scientists (TUBA GEBIP).

REFERENCES


