A Clustering-Based Approach for Filtering False Positive MicroRNA Candidates

W. S. Leung[1], Marie C. M. Lin[2], David W. Cheung[1], and S. M. Yiu[1]

[1] Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
[2] Department of Chemistry, Open Laboratory of Chemical Biology, The University of Hong Kong, Pokfulam Road, Hong Kong

Abstract. Our study first validated the phenomenon of microRNA (miRNA) clustering in the human genome using computational methods, and then showed that miRNA clustering can be used to improve the computational predictions of human miRNAs. We demonstrated that the secondary structure of a miRNA precursor is similar to its neighboring miRNAs located in the same cluster, when compared to the sequences outside the clusters. Using this property, we designed a 2-step approach to filter the false positives resulted from a miRNA software tool and successfully raised the specificity by 10% while keeping a reasonably high sensitivity.

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNA gene products of 19-25 nucleotides (nt) long, which function to repress the translation or mediate the degradation of their target messenger RNAs (mRNAs). A 22nt mature miRNA is derived from a precursor transcript of 60-80 nt in length, which is named as pre-miRNA. Pre-miRNAs can potentially fold into a hairpin structure without large internal loops or bulges. MiRNAs were found to play diversified roles from species to species [For reviews, see, e.g. [1, 2]]. In recent years, the increasing number of researches on the roles of miRNAs in cancers suggest that miRNAs may have important therapeutic potential in human diseases. To date, there are 541 human miRNA entries deposited in miRBase [3, 4], the home of miRNA data, in Release 10.1. Yet some studies suggested that the total number can reach at least 800 [5, 6], hence continual efforts should be made on locating the unknown miRNAs. A lot of computational prediction methods and tools have been developed over the years [For review, see [7]]. Here, we first validated the clustering phenomenon of human miRNA by computational means. We then applied this clustering concept to a selected software tool, ProMirII-g, aiming to filter its false positive miRNA predictions.

2 MiRNA Clustering

Many miRNAs are found to be arranged in clusters [8], meaning that they are in close proximity with their neighboring miRNAs. MiRNAs located in the same cluster are usually co-regulated and co-expressed, [9, 10], suggesting that miRNA clustering can be used to assist the prediction of novel miRNAs. In view of this, we analyzed how this idea could be applied computationally.

2.1 Analysis of MiRNA Clustering in the Human Genome

To assess the clustering property of miRNAs in the human genome, the genomic locations of all human miRNAs (Release 9.1) were studied. We defined that two miRNAs belong to the same cluster if the chromosomal distance between them is less than 3000nt. There are 52 clusters identified, which are equivalent to 40% of the total human miRNAs.

2.2 Similarity Analyses among Clustered MiRNAs, Non-Clustered MiRNAs, Neighboring Sequences of Clustered MiRNAs and Random Sequences

To assess the sequence and secondary structure similarities among miRNAs in the same cluster, each clustered miRNA was aligned with the sequences from the following four categories in a pairwise manner: (i) its fellow miRNAs found in the same cluster; (ii) miRNAs located outside its cluster; (iii) random sequences extracted from the genome; and (iv) neighboring sequences extracted from its flanking 3000nt regions. The software T-COFFEE [11] was used for pairwise sequence alignment. The program RNAdistance [12] was used to compute the distance between two miRNA secondary structures, which were determined by RNAfold. Table 1 summarizes the results of the sequence and secondary structure alignments. As shown in Table 1, there was no observable difference among

Table 1. Results of sequence and secondary structural alignments of clustered miRNAs with (i) clustered miRNAs, (ii) non-clustered miRNAs, (iii) random and (iv) neighboring sequences. A higher score implies a greater distance and hence a higher degree of dissimilarity.

	T-COFFEE sequence alignment scores				RNAdistance structure alignment scores			
Category	Maximum	Average	Minimum	Std Dev	Maximum	Average	Mininum	Std Dev
(i)	91	51.86	17	17.66	71	28.58	0	9.74
(ii)	87	49.57	18	17.34	70	36.35	11	10.28
(iii)	86	48.62	19	15.90	132	73.63	26	18.83
(iv)	89	48.60	19	16.06	121	72.60	32	19.10

the sequence alignment scores of the four categories, suggesting that sequence similarity is unlikely to be useful for identifying clustered miRNAs. Interestingly, the distance between the secondary structures of miRNAs located in the same cluster was found to be much smaller than the distances obtained by comparing the structures of clustered miRNAs with the other three categories. In other words, clustered miRNAs are structurally more similar to one another, and the RNAdistance score can be used to assess the structural similarity between two sequences. Based on this observation, we proposed a 2-step approach to improve the effectiveness of computational prediction of miRNAs.

3 Application of MiRNA Clustering: 2-Step Approach

3.1 Overview

We selected a software tool named ProMirII-g [13, 14] to test our proposed 2-step approach. In terms of specificity and sensitivity, we first analyzed the performance of ProMirII-g, which serves as a benchmark for comparison with our approach. Using a relaxed threshold of ProMirII-g, a high sensitivity and a low specificity are expected. The prediction results will consist of most true positives but a large number of false positives. Our approach aims at reducing as many false positives as possible with the application of miRNA clustering.

3.2 Performance Analysis of ProMirII-g

Methodology A total of 77 sequences of 10000nt-long were extracted from the human genome and served as the input sequences for ProMirII-g. Each input sequence consists of at least two miRNAs which are in close proximity with each other. There were four prediction thresholds provided by ProMirII-g. A larger threshold implies a stricter threshold. For each input sequence and each threshold, a list of the corresponding predicted candidates was resulted. The candidates were matched with the real miRNAs. Table 2 summarizes the specificity and sensitivity resulted from the four prediction thresholds.

Table 2. Performance analysis of ProMirII-g using human pre-miRNAs data. (Abbreviations: #, Number; FP, false positive; TP, true positive; SE, sensitivity; SP, specificity)

Prediction threshold	# of thresholds	# of TP	# of FP	# missed	SE	SP
0.001	690	215	475	26	89.21%	31.16%
0.017	269	184	85	57	76.35%	68.40%
0.033	220	165	55	76	68.46%	75.00%
0.33	104	84	20	155	34.85%	80.77%

3.3 Evaluation of the 2-Step Approach

Methodology In step 1, ProMirII-g with prediction threshold set at 0.001 was run with the same set of input sequences as described above. A list of miRNA candidates was generated. Recall that each input sequence consists of miRNAs located in the same cluster, the candidates were thus potential clustered miRNAs. In step 2, pairwise structural alignment between each pair of candidates was conducted using RNAdistance. Since a higher RNAdistance score implies that the two candidate sequences have relatively different structures and vice versa, if a candidate has high pairwise RNAdistance scores with other candidates, it is likely to be a false positive and should be eliminated. An adjustable threshold is needed to determine candidate(s) with high scores.

Results We conducted several preliminary tests of our 2-step approach using different thresholds and selected the best result as the new set of predictions. We compared it with the real miRNAs and obtained 42.86% and 85.89% for specificity and sensitivity respectively. With reference to the benchmark reported above, our 2-step approach successfully raised the specificity by more than 10% while keeping a reasonably high sensitivity.

4 Conclusion

We validated the phenomenon of miRNA clustering in the human genome and demonstrated that miRNAs located in the same cluster are structurally similar to one another. We then applied the structural similarity property of clustered miRNAs to propose a 2-step approach to improve miRNA prediction. We tested our approach on the software ProMirII-g, and were able to raise its specificity by 10% without significant change in sensitivity. More tests will be conducted to further verify the effectiveness of this 2-step approach. Data from different mammalian genomes and different miRNA prediction software tools will be used.

References

1. Ambros, V.: The functions of animal microRNAs. Nature **431** (2004) 350–355
2. Cullen, B.R.: Viruses and microRNAs. Nat Genet. **38** (2006) S25–S30
3. Griffiths-Jones, S.: The microRNA Registry. Nucleic Acids Res. **32** (2004) D109–D111
4. Griffiths-Jones, S., Grocock, R.J., et al.: miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. **34** (2006) D140–D144
5. Bentwich, I., Avniel, A., et al.: Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet **37** (2005) 766–770
6. Berezikov, E., Guryev, V., et al.: Phylogenetic Shadowing and Computational Identification of Human microRNA Genes. Cell **120** (2005) 21–24
7. Leung, W.S., Yiu, S.M., et al.: Computational prediction on mammalian and viral microRNAs - A review. IJIB **1** (2007) 118–126
8. Altuvia, Y., Landgraf, P., et al.: Clustering and conservation patterns of human microRNAs. Nucleic Acids Res. **33** (2005) 2697–2706
9. Tanzer, A., Stadler, P.F.: Molecular evolution of a microRNA cluster. J Mol Biol. **339** (2004) 327–335
10. Tanzer, A., Stadler, P.F.: Evolution of microRNAs. Methods Mol Biol. **342** (2006) 335–350
11. Notredame, C., Higgins, D., et al.: T-Coffee: A novel method for multiple sequence alignments. J Mol Biol. **302** (2000) 205–217
12. Hofacker, I.L.: Vienna RNA secondary structure server. Nucleic Acids Res. **31** (2003) 3429–3431
13. Nam, J.W., Shin, K.R., et al.: Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res. **33** (2005) 3570–3581
14. Nam, J.W., Kim, J., et al.: ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res. **34** (2006) W455–W458