Evaluation of a Kernel Function for Recognizing microRNAs

Lidio M. C. Meireles
Joint CMU-Pitt PhD Program in Computational Biology
Department of Computational Biology
University of Pittsburgh
lmm85@pitt.edu

1 Introduction

MicroRNA (miRNA) is a small non-coding RNA molecule of about 21 nucleotides that regulates gene expression at the post-transcriptional level. The miRNA silences target genes by making complementary base pairs with the gene's messenger RNAs, leading to their degradation or translational repression. The biogenesis and function of miRNAs have been reviewed in [1, 6].

MicroRNAs were recently discovered, in 1993, when it was first reported that a small RNA in *Caenorhabditis elegans*, called *lin-4*, was responsible for regulating the expression of the *lin-14* gene through direct interaction with its messenger RNA [8, 13]. A few years later, Andrew Z. Fire and Craig C. Mello published a paper in *Nature* [3] describing how tiny snippets of RNA can destroy the gene's messenger RNA before it can produce a protein. Scientists then started to explore this mechanism, named *RNA interference* (RNAi), to silence genes of therapeutic interests. RNAi has become one of the most important recent developments in molecular biology, as exemplified by the fact that Fire and Mello were awarded this year Nobel Prize in Medicine [15] for their discovery.

The contribution of computational biology to miRNA research are threefold [2]: (1) identification of new microRNA genes in genomes; (2) prediction of microRNA gene targets; and (3) computational design of microRNAs to target therapeutic genes. In this report, we consider only the first problem, i.e., the problem of automatic recognition of microRNA genes in genomic sequences.

2 Problem Definition

A *microRNA precursor (pre-miRNA)* is a RNA sequence of about 100 nucleotides that contains the actual microRNA (of $\sim$ 21 nt, also called *mature miRNA*). The mature miRNA is cleaved from its precursor by specific enzymes. The miRNA recognition problem is usually defined over pre-miRNAs because they encode more information (to be exploited by recognition algorithms) than the smaller mature miRNA. In particular, the pre-miRNAs have a typical hairpin loop secondary structure as shown in Fig. 1.

Here we define the recognition problem as follows. Given a RNA sequence of $\sim$ 100 nt, determine whether it is a miRNA or not. The problem is more interesting (and harder!) at those cases where the input RNA sequence folds like a typical hairpin loop of miRNA, so

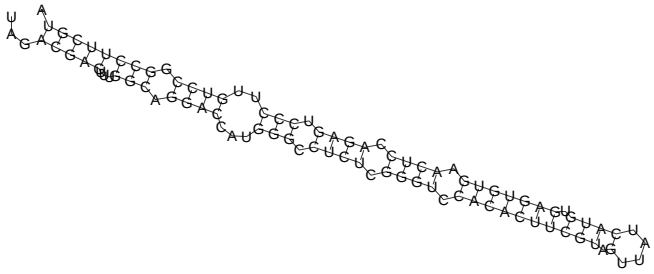

Figure 1: Secondary structure of *C. elegans* lin-4 miRNA folded by RNAfold [5].

the algorithm should be able to discriminate between true miRNA hairpin loops from all other hairpin loops.

3 Existing Methods

The simplest way to identify new miRNAs is through sequence homology searches using programs, such as BLAST. However, because miRNAs often have non-conserved sequences, this approach is limited to finding only a small fraction of miRNAs with close homologs. Other approaches explore the fact that RNA secondary structures tends to be more conserved than their sequences. Such approaches rely on programs, like RNAfold [5], to predict the RNA secondary structure. The predicted structures are then used to identify potential miRNA candidates that fold like a typical miRNA loop. We are particularly interested in machine learning approaches applied to this problem.

Currently, there are three machine learning approaches applied to miRNA identification found in the literature, namely, support vector machines (SVM), Naive Bayes and hidden Markov model (HMM). In the SVM [10] and Naive Bayes [14] approaches, a feature vector is extracted from a combination of sequence and (predicted) structure. Some features include the composition of bases in specific positions, the number and size of bulges, base pairing mismatches, and so forth. The HMM approach [12] has a topology of hidden states designed to learn the base-pairing compositions along the miRNA hairpin-loop. It is difficult to say which approach is the most effective; since they were trained using different datasets, their reported results are not directly comparable.

4 New Proposed Method

Since the existing approaches rely on predicted structures, we reason that their accuracy are somehow limited by the accuracy of RNA secondary structure prediction programs. While the overall hairpin loop structure is predicted well, fine details like the size of loops and bulges are not correctly predicted, resulting in noisy feature vectors. In [7], it was reported that 8 structures of miRNAs out of 10 were incorrectly predicted. Therefore, we propose a new machine learning approach that accounts for the secondary structure in an implicit way without relying on a specific predicted structure, through the use of a kernel function that computes a similarity measure for two RNA sequences.

4.1 New Kernel Function

We have designed a new kernel function that computes a similarity measure for two RNA sequences based on their patterns of base-pairing formation. We shall describe the kernel function in two steps: (1) computation of a *base-pair profile* for each RNA sequence, and (2) alignmnent of the two base-pair profiles.

The base-pair profile of a RNA molecule should capture its pattern of base-pairing formation. Rather than relying on a single predicted structure, we use the McCaskill algorithm [11] to compute base-pair probabilities based on thermodynamics principles. The result is a matrix $Pr[i, j]$ where each entry stores the probability of base i forming a pair with base j. The base-pair profile is a Nx3 matrix $PROFILE$ (where N is the sequence length) that gives you for each base i its probability of forming base pairing upstream (US) , downstream (DS) and not forming base pairing (NP). These are readily computed from $Pr[i, j]$ as follows:

$$PROFILE[i, UP] = \sum_{j>i} Pr[i, j] \tag{1}$$

$$PROFILE[i, DS] = \sum_{j<i} Pr[i, j] \tag{2}$$

$$PROFILE[i, NP] = 1 - PROFILE[i, UP] - PROFILE[i, DS]; \tag{3}$$

The second step is the global alignment of two $PROFILE$'s. The alignment is computed using the Needlman-Wunch algorithm with a modified scoring system. We allow gaps with zero cost and we score the alignment of two bases by the inner product of their probability profile vectors. See the recurrence equation below.

$$\text{Score[u,v]} = \max \begin{cases} \text{Score[u-1,v]} & \text{/* zero gap cost */} \\ \text{Score[u,v-1]} & \text{/* zero gap cost */} \\ \text{Score[u-1,v-1]} + \sum_{i=1..3} PROFILE_1[u, i] * PROFILE_2[v, i] & \text{/* inner product */} \end{cases} \tag{4}$$

The kernel function output is the score of the best alignment of base-pair profiles of the given RNA sequences. It should be clear that this kernel function returns higher values for RNA sequences that shares a common pattern of base-pairing formation.

5 Experiments

We performed two experiments with different sets of negative examples. In the first experiment, we used negative random sequences, and in the second experiment, we used

segments of messenger RNA from *C. elegans* that were predicted to fold like a hairpin loop by the program SRNAloop [4]. In both experiments, we used 79 miRNA sequences from *C. briggsae* as positive examples [16]. The ratio of positive to negative examples was 1/1. We used the SVM package libSVM and 4-fold cross validation. The results are given below.

Experiment 1: 93.7 % accuracy.

Experiment 2: 89.8 % accuracy.

6 Conclusion

We claim our method is very promising since its accuracy of 89.8% was obtained using a small training set and without performing any parameter optimization (due to lack of time). Probably even better results could be obtained by optimizing the SVM parameters and the alignment score function (e.g. gap cost). Moreover, comparing with other methods in the literature our method has the advantage of not being miRNA specific, i.e. it can also be applied to classify among different families of RNA's.

References

[1] Bartel, D.P. (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function *Cell*, **116**, 281–297.

[2] Brown, J.R., Sanseau, P. (2005) A computational view of microRNAs and their targets. *DDT:Biosilico*, **10**:8, 595–601.

[3] Fire, A.Z. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.

[4] Grad, Y. *et al.* (2003) Computational and Experimental Identification of *C. elegans* microRNAs. *Molecular Cell*, **11**:5, 1253–1263.

[5] Hofacker, I.L., (2003) Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**:13, 3429–3431.

[6] Kim, V.N., Nam, J.W. (2006) Genomics of microRNA *TRENDS in Genetics*, **22**:3, 165–173.

[7] Krol, J. *et al.* (2004) Structural features of MicroRNA (miRNA) Precursors and Their Relevance to miRNA Biogenesis and Small Interfering RNA/Short Hairpin RNA Design, *The Journal of Biological Chemistry*, **279**:40, 42230–39.

[8] Lee, R.C. *et al.* (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

[9] Leslie, C. *et al.* (2002) The spectrum kernel: A string kernel for SVM protein classification. *Proc. of the Pacific Symposium on Biocomputing*.

[10] Liang Huai Yang, Wynne Hsu, Mong Li Lee, Limsoon Wong. (2006) SVM-based Identification of microRNA Precursors, *Proceedings of 4th Asia-Pacific Bioinformatics Conference*, 267–276, Taipei, Taiwan, February 2006.

[11] McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**:(6-7), 1105–19.

[12] Nam, J.W., *et al.* (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*, **33**:11, 3570–3581.

[13] Wightman, B. *et al.* (1993) Post-transcriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, **75**, 855–862.

[14] Yousef, M. *et al.* (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier, *Bioinformatics*, **22**:11, 1325–1334.

[15] http://nobelprize.org/nobel_prizes/medicine/laureates/2006/

[16] miRBase Sequence Database. http://microrna.sanger.ac.uk/