Mature microRNA identification via the use of a Naive Bayes classifier

Master Thesis

Gkirtzou Katerina

Computer Science Department
University of Crete

13/03/2009

Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

MicroRNAs

Definition

MicroRNAs (miRNAs) are small single stranded RNAs, on average 22nt long, generated from endogenous hairpin–shaped transcripts with post transcriptional activity.

MicroRNAs

Definition

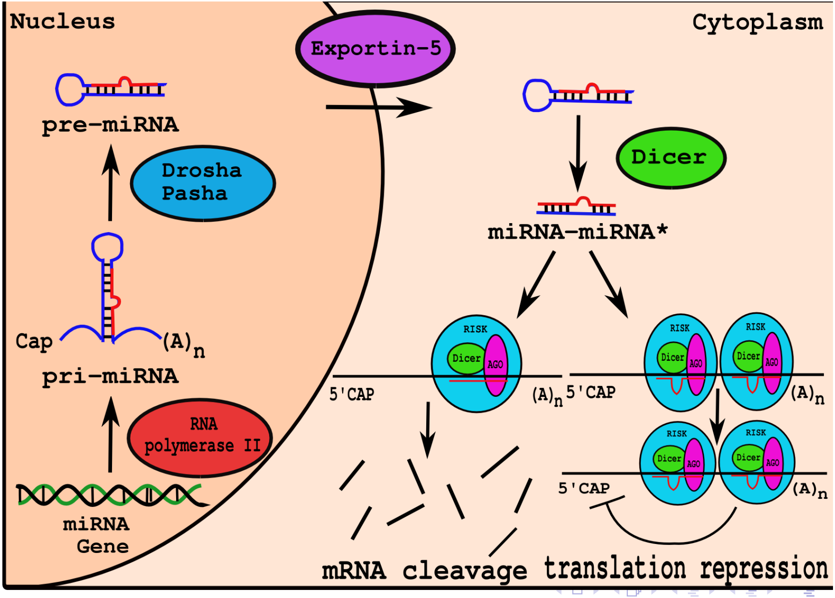
MicroRNAs (miRNAs) are small single stranded RNAs, on average 22nt long, generated from endogenous hairpin–shaped transcripts with post transcriptional activity.

Significance

MicroRNAs have been observed to participate in:

- Developmental timing.
- Cell proliferation and cell differentiation.
- Apoptosis.
- Diseases, such as diabetes and cancer.
- Anti–viral defense.

MicroRNA Biogenesis



Related Work

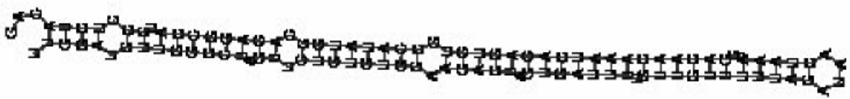
Disadvantages

- Hypothesize that only a single mature is produced from every hairpin structure (BayesMiRNAfind – 2006, ProMiR – 2006)
- Hypothesize that pri-miRNAs are always processed by the Drosha complex, whose cleavage cite determines the start position of the mature (Microprocessor SVM – 2006).
- Mature candidate is provided only for the human precursors, which are expressed in specific cell lines (SSCprofiler – 2009).
- Evaluation of performance is often measured in terms of true positive rate alone, ignoring the false positive rate (ProMiR – 2006, BayesMiRNAfind – 2006, Tao – 2007, mirCoS – 2007, SSCprofiler – 2009)
- Distance distribution of predicted compared to true matures is not provided (BayesMiRNAfind – 2006, Tao – 2007, mirCoS – 2007, SSCprofiler – 2009).

Objectives

Goal

Build a classifier considering biological information of precursor miRNA, such as sequence or secondary structure, capable of identifying the mature miRNA(s) within a precursor miRNA with high accuracy.



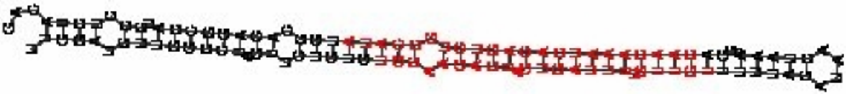
Objectives

Goal

Build a classifier considering biological information of precursor miRNA, such as sequence or secondary structure, capable of identifying the mature miRNA(s) within a precursor miRNA with high accuracy.

Output

The model's output is the predicted *start position* of the mature miRNA(s) for each precursor sequence.



Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

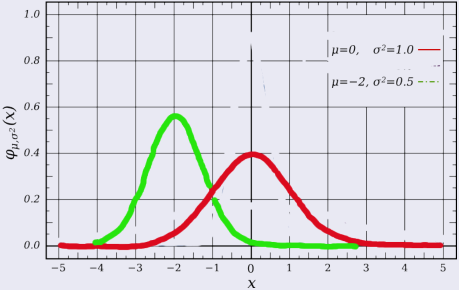
4. Conclusions

Naive Bayes Classifier

Advantages of Naive Bayes

- Achieves strong performance in many real problems, despite its simplified assumptions.
- Requires small amount of training data.
- Contribution of features is easily derived.

Our decision surface



$$\frac{P(C_{mature}|\mathbf{x})}{P(C_{non-mature}|\mathbf{x})} > \lambda$$

Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

Datasets

Typical two class classification problem

- Positive data: experimental verified mature miRNAs.
- Negative data: What is a non–mature miRNA?

Datasets

Typical two class classification problem

- Positive data: experimental verified mature miRNAs.
- Negative data: What is a non–mature miRNA?

Observations

- Known miRNA precursors do not produce multiple overlapping mature miRNAs from the same arm of the foldback precursor.
- The search area of the classifier will be a miRNA precursor sequence.

Datasets

Typical two class classification problem

- Positive data: experimental verified mature miRNAs.
- Negative data: What is a non–mature miRNA?

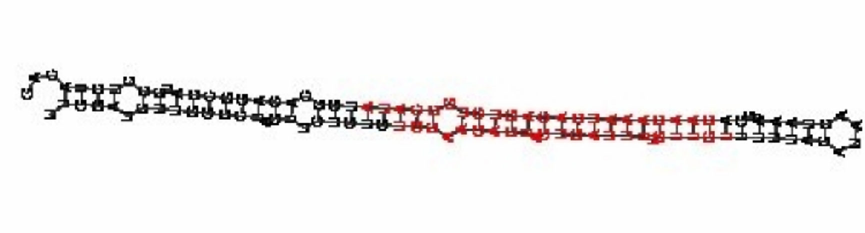
Observations

- Known miRNA precursors do not produce multiple overlapping mature miRNAs from the same arm of the foldback precursor.
- The search area of the classifier will be a miRNA precursor sequence.

Solution!

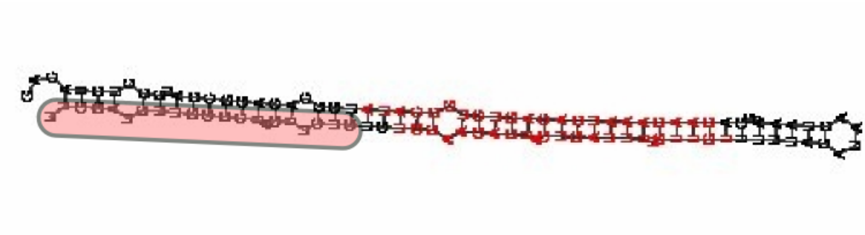
Create negative data by sliding 1 base pair in both stem arms of the precursor with a window with size the same as the produced mature miRNA.

Produce Negative Data



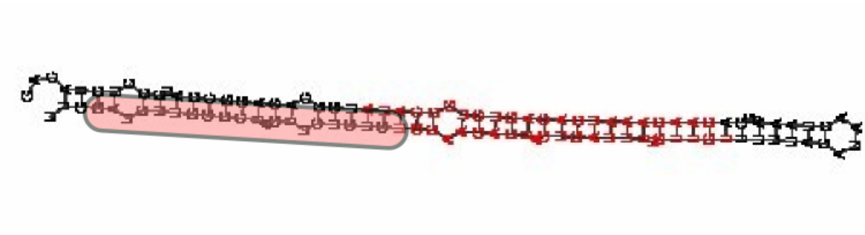
Negative Data		Positive Data	
SP	EP	SP	EP

Produce Negative Data



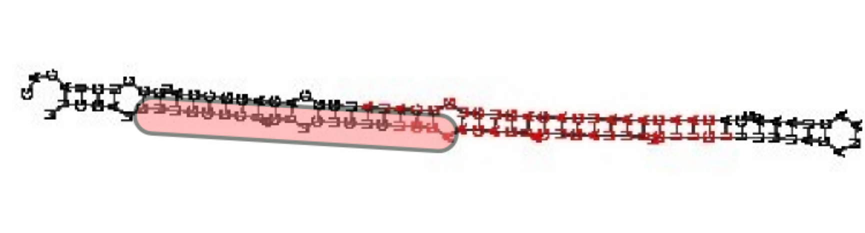
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		

Produce Negative Data



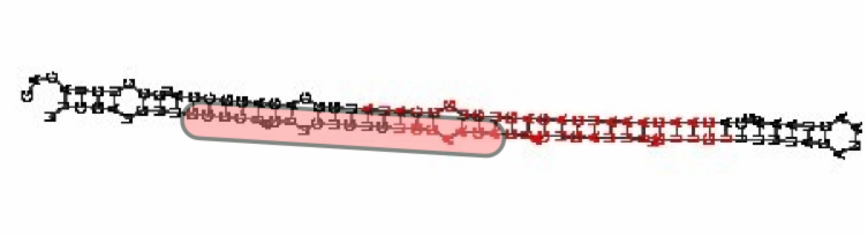
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		
2	23		

Produce Negative Data



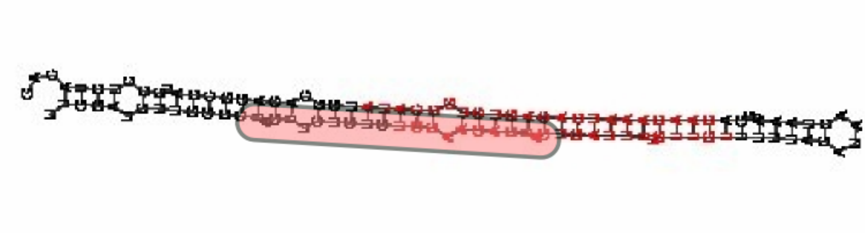
Negative Data		Positive Data	
SP	**EP**	**SP**	**EP**
1	22		
2	23		
3	24		

Produce Negative Data



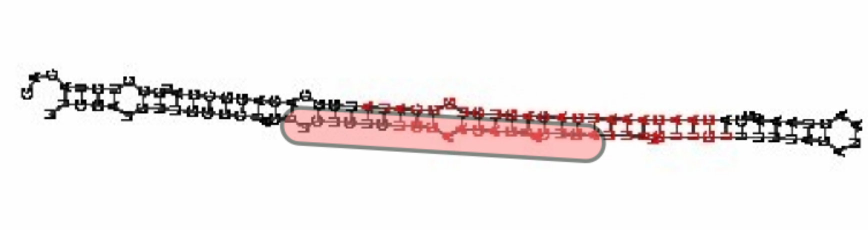
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		
2	23		
3	24		
⋮	⋮		

Produce Negative Data



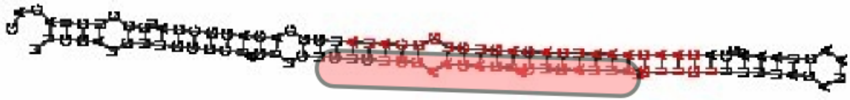
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		
2	23		
3	24		
⋮	⋮		

Produce Negative Data



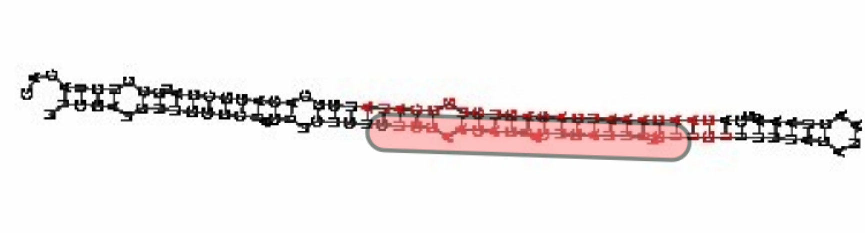
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		
2	23		
3	24		
⋮	⋮		

Produce Negative Data



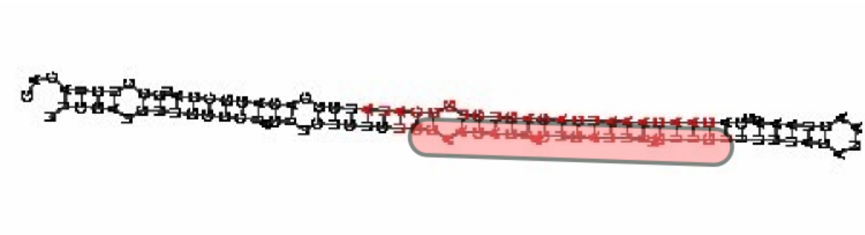
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		
2	23		
3	24		
⋮	⋮		

Produce Negative Data



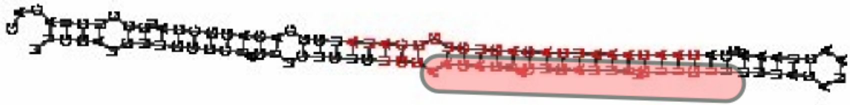
Negative Data		Positive Data	
SP	EP	SP	EP
1	22		
2	23		
3	24		
⋮	⋮		

Produce Negative Data



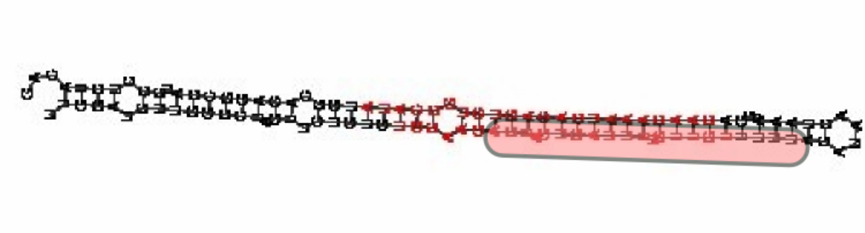
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23		
3	24		
⋮	⋮		

Produce Negative Data



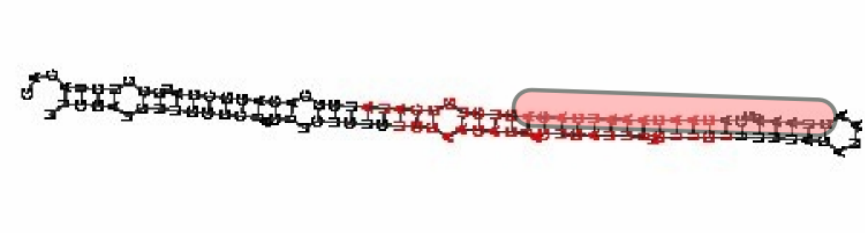
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23		
3	24		
⋮	⋮		

Produce Negative Data



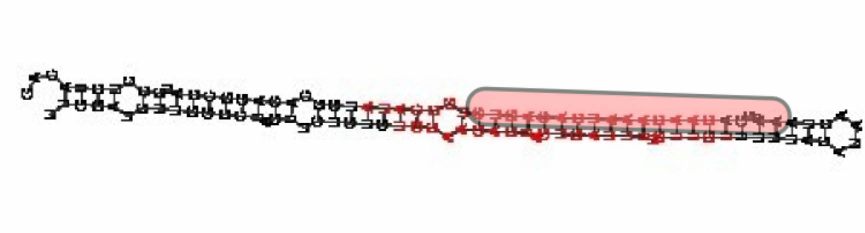
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23		
3	24		
⋮	⋮		

Produce Negative Data



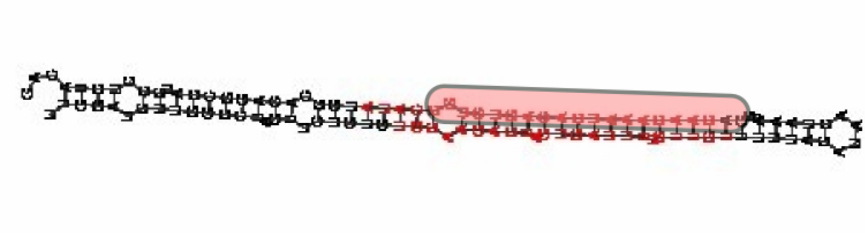
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23		
3	24		
⋮	⋮		

Produce Negative Data



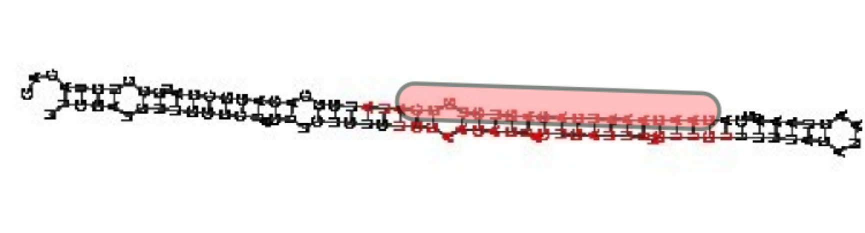
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23		
3	24		
⋮	⋮		

Produce Negative Data



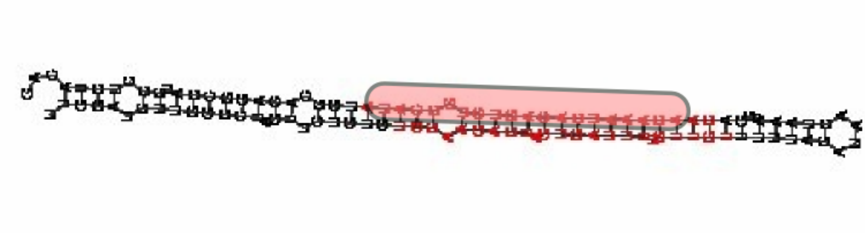
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23		
3	24		
⋮	⋮		

Produce Negative Data



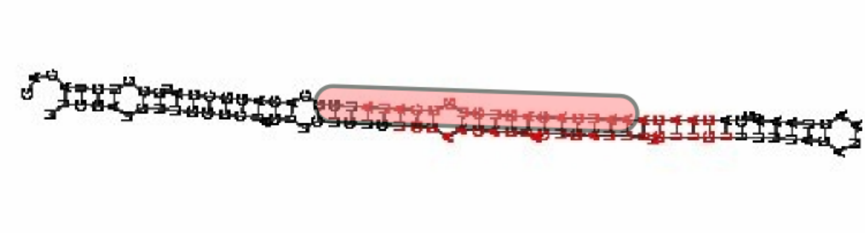
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		

Produce Negative Data



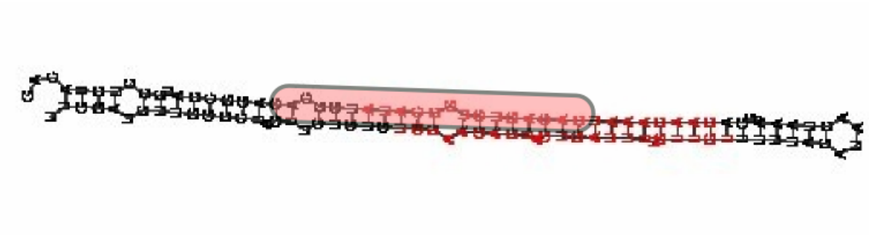
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		

Produce Negative Data



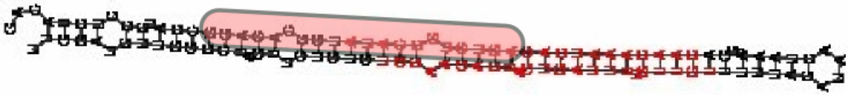
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		

Produce Negative Data



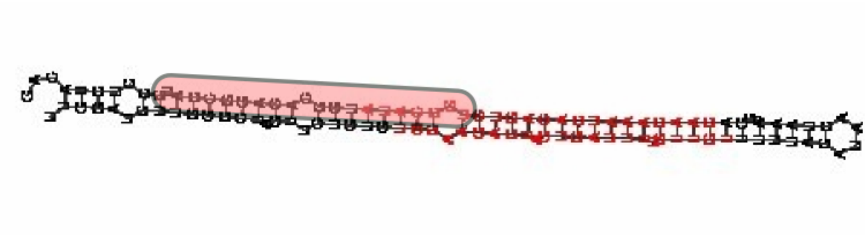
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		

Produce Negative Data



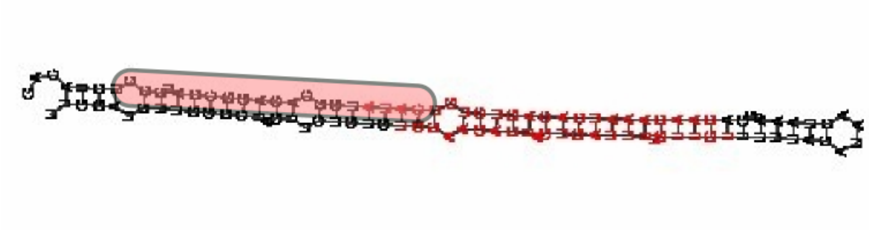
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		

Produce Negative Data



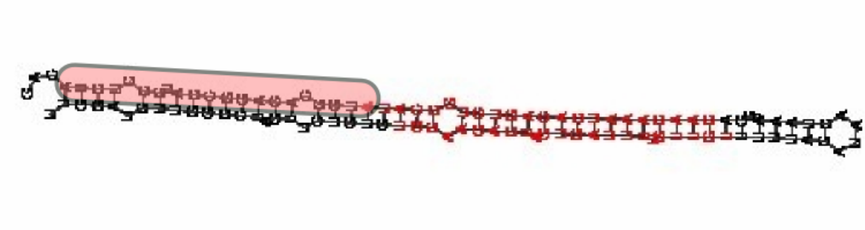
Negative Data		Positive Data	
SP	**EP**	**SP**	**EP**
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		

Produce Negative Data



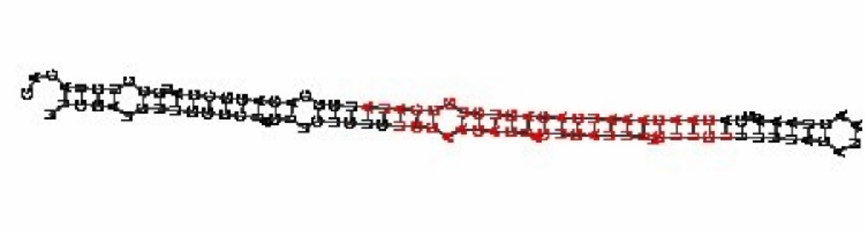
Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		
59	81		

Produce Negative Data



Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		
59	81		
60	82		

Produce Negative Data



Negative Data		Positive Data	
SP	EP	SP	EP
1	22	24	46
2	23	40	62
3	24		
⋮	⋮		
59	81		
60	82		

Outline

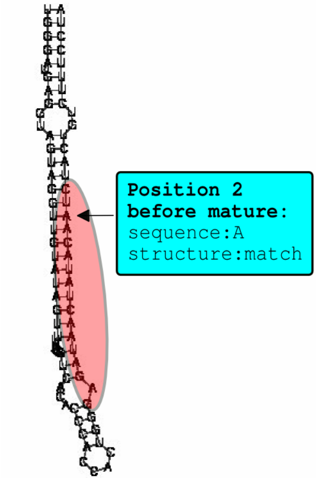
1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

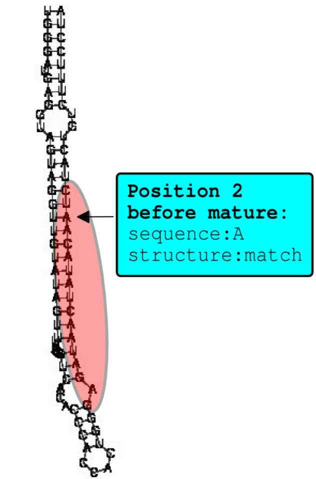
Position oriented features



Position 2
before mature:
sequence:A
structure:match

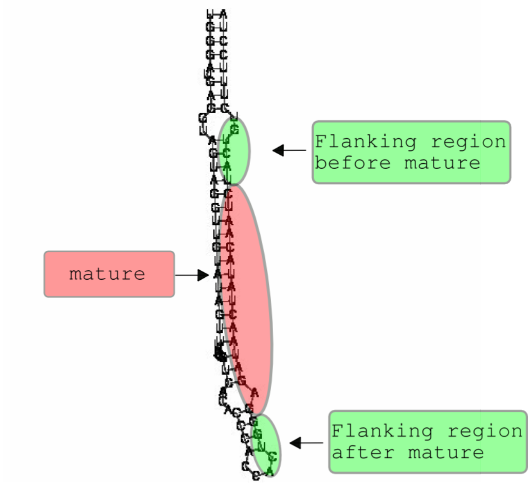
Example of a position oriented features

Position oriented features



Position 2 before mature:
sequence:A
structure:match

Example of a position oriented features



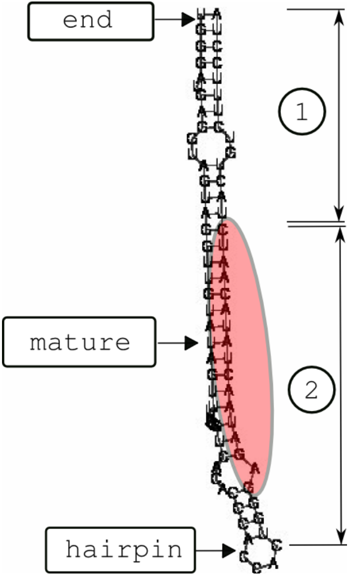
mature

Flanking region before mature

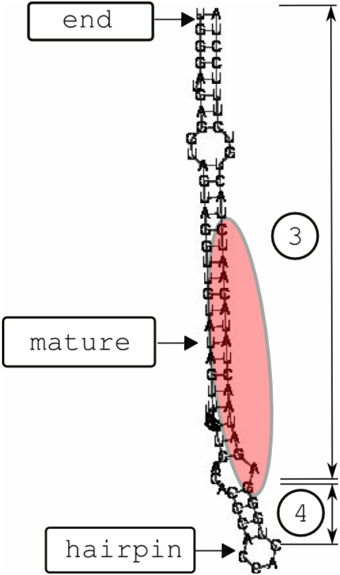
Flanking region after mature

Areas of position oriented features

Distance oriented features



Distance of Starting Position



Distance of Ending Position

Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

Feature Selection

Feature Selection Ranking Method

1. For each feature estimate the probability mass functions in both positive and negative data.

Feature Selection

Feature Selection Ranking Method

1. For each feature estimate the probability mass functions in both positive and negative data.
2. Using the symmetric K-L divergence estimate a score for each feature.

Symmetric Kullback–Leibler divergence

The divergence between the positive (P) and negative (N) probability distribution:

$$Sym\_D_{KL} = \frac{1}{2}\left(D_{KL}(P||N) + D_{KL}(N||P)\right)$$

where $D_{KL}(P||N) = \sum_i P(i) \log_2 \frac{P(i)}{N(i)}$

Feature Selection

Feature Selection Ranking Method

1. For each feature estimate the probability mass functions in both positive and negative data.
2. Using the symmetric K-L divergence estimate a score for each feature.
3. Rank features according to the K-L provided score.

Symmetric Kullback–Leibler divergence

The divergence between the positive (P) and negative (N) probability distribution:

$$Sym\_D_{KL} = \frac{1}{2}\left(D_{KL}(P||N) + D_{KL}(N||P)\right)$$

where $D_{KL}(P||N) = \sum_i P(i) \log_2 \frac{P(i)}{N(i)}$

Feature Selection

Feature Selection Ranking Method

1. For each feature estimate the probability mass functions in both positive and negative data.
2. Using the symmetric K-L divergence estimate a score for each feature.
3. Rank features according to the K-L provided score.
4. Train the classifier using the top K features. Incoporate features gradually only if it helps increasing the performance of the classifier.

Symmetric Kullback–Leibler divergence

The divergence between the positive (P) and negative (N) probability distribution:

$$Sym\_D_{KL} = \frac{1}{2}\left(D_{KL}(P||N) + D_{KL}(N||P)\right)$$

where $D_{KL}(P||N) = \sum_i P(i) \log_2 \frac{P(i)}{N(i)}$

Outline

1. Introduction
2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection
3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods
4. Conclusions

Datasets

Training Dataset – Version 10.1 miRBase

Organism	Precursor	True Mature	Negative Mature
Human	533	729	7290
Mouse	422	530	5300

Test Dataset – Version 12 miRBase

Organism	Precursor	True Mature
Human	155	160
Mouse	45	48

Implementation Specifications

Extra Parameters: tune over a 10–fold cross validation

- The size of the flanking regions, N.
- The size of the scanning window, W.
- The number of features used in the classifier, K .
- The type of information the *position oriented* features hold.

Implementation Specifications

Extra Parameters: tune over a 10–fold cross validation

- The size of the flanking regions, *N*.
- The size of the scanning window, *W*.
- The number of features used in the classifier, *K* .
- The type of information the *position oriented* features hold.

Evaluation Specification

- The validation sets consisted of true miRNA precursors, whose mature miRNAs were left out from training in the cross validation procedure.
- Candidates mature miRNAs were produced by sliding 1 base pair in both stem arms of the precursor with a fixed size sliding window, *W*.
- Evaluation was estimated based on exact match of the starting position of the predicted compared to the real mature miRNA.

Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

Information contained in Position Oriented Features

Classifier's Description	Sensitivity	Specificity
Sequence Based Naive Bayes Classifiers		
0nt flanking region	67.10%	55.10%
5nt flanking region	76.04%	53.34%
7nt flanking region	75.96%	53.20%
10nt flanking region	79.15%	47.01%
12nt flanking region	74.30%	51.33%
Structure Based Naive Bayes Classifiers		
0nt flanking region	65.70%	54.30%
5nt flanking region	76.34%	52.64%
7nt flanking region	77.85%	54.29%
10nt flanking region	81.01%	56.63%
12nt flanking region	79.89%	55.51%
Combined Naive Bayes Classifiers		
0nt flanking region	68.50%	62.50%
5nt flanking region	71.32%	65.34%
7nt flanking region	74.26%	66.46%
10nt flanking region	76.50%	65.61%
12nt flanking region	77.81%	64.14%

Information contained in Distance Oriented Features

Distance Oriented Naive Bayes Classifiers – AUC

Distance oriented Features	**Window 18nt**	**Window 20nt**	**Window 22nt**	**Window 24nt**
HS	0.8181	0.8155	0.8128	0.8147
HS-HE	0.7794	0.7914	0.8099	0.8100
HS-HE-ES	0.7621	0.7803	0.7787	0.7866
HS-HE-ES-EE	0.7587	0.7808	0.7875	0.7839

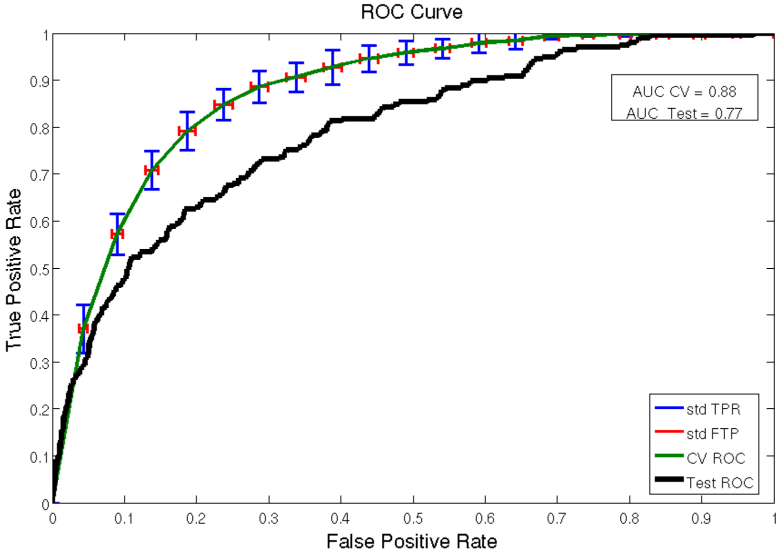
- **HS** : the distance of the **starting** position of the mature miRNA from the **hairpin**.
- **HE** : the distance of the **ending** position of the mature miRNA from the **hairpin**.
- **ES** : the distance of the **starting** position of the mature miRNA from the **ends** of the precursor.
- **EE** : the distance of the **ending** position of the mature miRNA from the **ends** of the precursor.

Searching the Optimun Naive Bayes Classifier

HS and Position Oriented Naive Bayes Classifiers – AUC

Flanking Region	Window 18nt	Window 20nt	Window 22nt	Window 24nt
0nt	0.8629	0.8615	0.8621	0.8624
3nt	0.8671	0.8658	0.8675	0.8661
5nt	0.8597	0.8614	0.8662	0.8642
7nt	0.8592	0.8630	0.8716	0.8696
9nt	0.8599	0.8673	**0.8771**	0.8704
12nt	0.8585	0.8691	0.8745	0.8658

The ROC Curve of the Best Naive Bayes Classifier

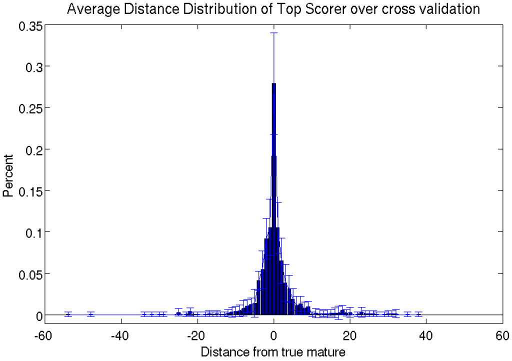


Outline

1. Introduction
2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection
3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods
4. Conclusions

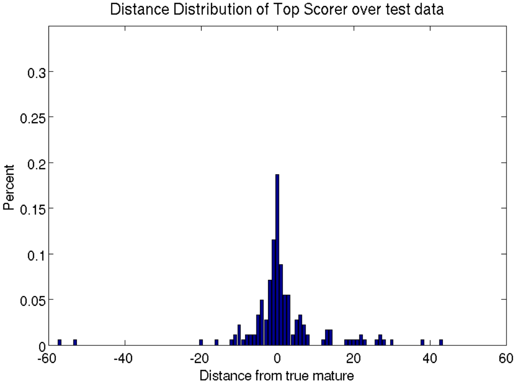
Top Scorer in Cross Validation

Distance from truth	Percent
0	**27.89%**
±1	48.91%
±2	64.59%
±3	73.92%
±4	**81.18%**
±5	84.48%
±6	86.88%
±7	89.28%

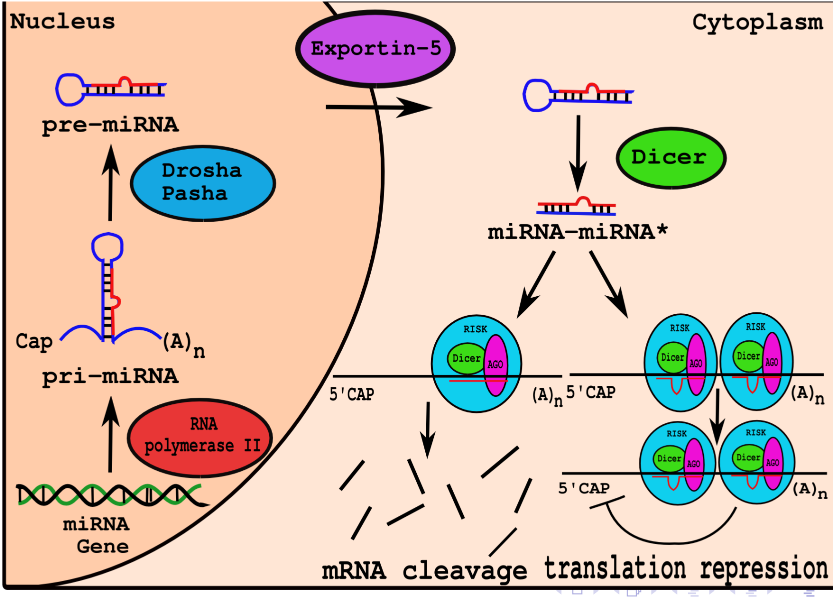


Top Scorer in Test Data

Distance from truth	Percent
0	**18.68%**
±1	39.01%
±2	51.61%
±3	59.89%
±4	**65.93%**
±5	71.98%
±6	76.37%
±7	79.67%

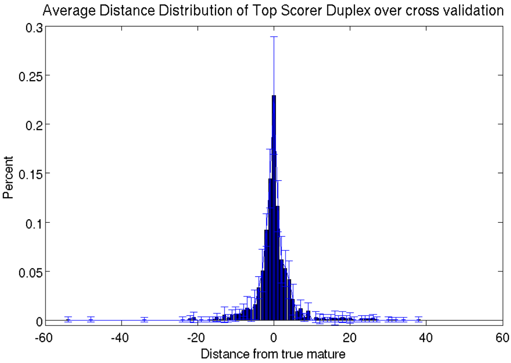


MicroRNA Biogenesis



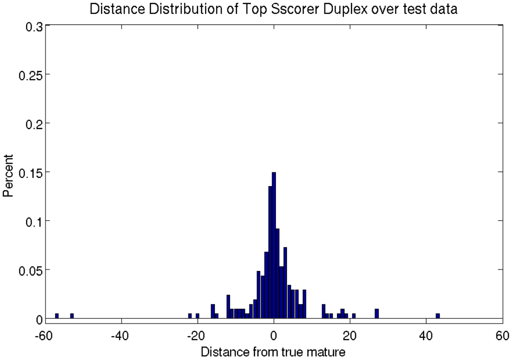
Top Scorer and its Duplex in Cross Validation

Distance from truth	Percent
0	**22.89%**
±1	48.97%
±2	64.35%
±3	74.71%
±4	**82.17%**
±5	85.87%
±6	87.83%
±7	90.30%



Top Scorer and its Duplex in Test Data

Distance from truth	**Percent**
0	**14.98%**
±1	37.68%
±2	49.76%
±3	61.35%
±4	**69.57%**
±5	74.40%
±6	78.74%
±7	80.68%



Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

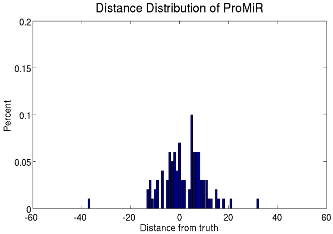
Comparison with ProMiR – Dataset Analysis

- Initial Dataset: 200 experimental human and mouse precursors.
- ProMiR predicted as precursors: 178/200.
- ProMiR predicted wrong stem for 78/178.
- Our Model predicted wrong stem for 94/178 if we consider as computational truth the top scorer of the precursor.
- Our Model predicted wrong stem for 0/178 if we consider as computational truth the top scorer of the precursor and its duplex.

Distance Distributions for Correct Stem Prediction

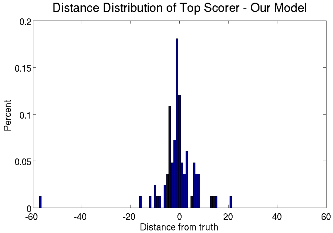
ProMiR

Distance from truth	Percent
0	**7%**
±1	12%
±2	23%
±3	28%
±4	**36%**
±5	49%
±6	55%
±7	65%



Top Scorer of our Model

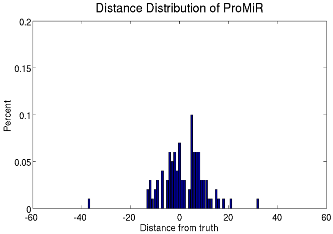
Distance from truth	Percent
0	**12.05%**
±1	34.94%
±2	45.78%
±3	56.63%
±4	**67.47%**
±5	72.29%
±6	79.52%
±7	83.13%



Distance Distributions for Correct Stem Prediction

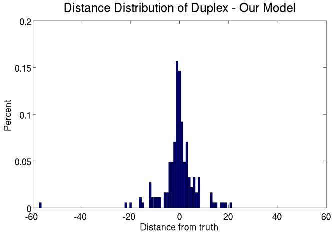
ProMiR

Distance from truth	Percent
0	**7%**
±1	12%
±2	23%
±3	28%
±4	**36%**
±5	49%
±6	55%
±7	65%


Distance Distribution of ProMiR

Top Scorer and its Duplex of our Model

Distance from truth	Percent
0	**14.59%**
±1	39.46%
±2	51.35%
±3	63.24%
±4	**71.35%**
±5	75.14%
±6	80.00%
±7	81.62%


Distance Distribution of Duplex - Our Model

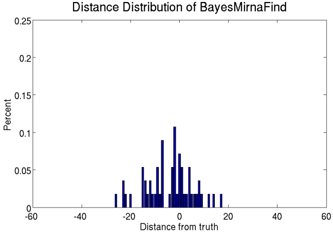
Comparison with BayesMiRNAfind – Dataset Analysis

- Initial Dataset: 200 experimental human and mouse precursors.
- BayesMiRNAfind predicted as precursors: 101/200.
- BayesMiRNAfind predicted wrong stem for 45/101.
- Our Model predicted wrong stem for 53/101 if we consider as computational truth the top scorer of the precursor.
- Our Model predicted wrong stem for 0/101 if we consider as computational truth the top scorer of the precursor and its duplex.

Distance Distributions for Correct Stem Prediction

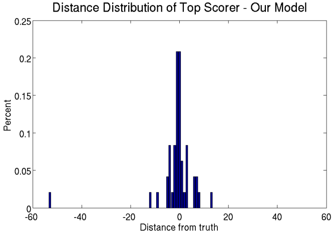
BayesMiRNAfind

Distance from truth	Percent
0	**7.14%**
±1	14.29%
±2	26.79%
±3	33.93%
±4	**41.07%**
±5	42.86%
±6	44.64%
±7	55.36%



Top Scorer of our Model

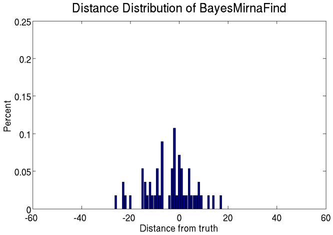
Distance from truth	Percent
0	**20.83%**
±1	47.92%
±2	58.33%
±3	68.75%
±4	**77.08%**
±5	81.25%
±6	85.42%
±7	89.58%



Distance Distributions for Correct Stem Prediction

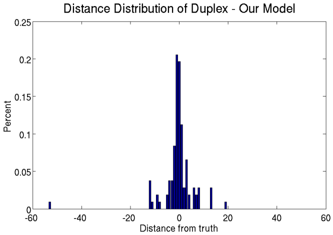
BayesMiRNAfind

Distance from truth	Percent
0	**7.14%**
±1	14.29%
±2	26.79%
±3	33.93%
±4	**41.07%**
±5	42.86%
±6	44.64%
±7	55.36%



Top Scorer and its Duplex of our Model

Distance from truth	Percent
0	**19.63%**
±1	51.40%
±2	62.62%
±3	72.90%
±4	**78.50%**
±5	80.37%
±6	83.18%
±7	85.05%



Outline

1. Introduction

2. Methodology
   - Naive Bayes Classifier
   - Datasets
   - Feature Definition
   - Feature Selection

3. Results
   - Training the Naive Bayes Classifier
   - Finding the Best Mature Candidate
   - Comparison with Other Methods

4. Conclusions

Conclusions

Innovations

- Feature Selection is based on Kullback–Leibler divergence.
- Performance is estimated based on AUC, in comparison with other methods that their performance are estimated based on sensitivity.
- Provide distance distributions for true matures.
- Flexibility to select between top scorer per stem or top scorer and its duplex per precursor.
- Simple algorithm with quite strong performance.

Conclusions

Innovations

- Feature Selection is based on Kullback–Leibler divergence.
- Performance is estimated based on AUC, in comparison with other methods that their performance are estimated based on sensitivity.
- Provide distance distributions for true matures.
- Flexibility to select between top scorer per stem or top scorer and its duplex per precursor.
- Simple algorithm with quite strong performance.

Comparison

Program	**Percent for ±4nt**	**Program**	**Percent for ±4nt**
ProMir	36.00%	BayesMiRNA	41.07%
Top Scorer	67.47%	Top Scorer	77.08%
Duplex	71.35%	Duplex	78.50%

Conclusions

Conclusion

Our findings suggest that position specific sequence and structure information and the distance of the starting position from the hairpin combined with a simple Bayes classifier achieve a good performance on the challenging task of mature miRNA identification.

Future Work

- Examine different error costs per class.
- Use stronger classifier, such as support vector machines (SVM).
- Use as training input the miRNA–miRNA* duplex.

Publications

K. Gkirtzou, P. Tsakalides and P. Poirazi.
Mature microRNA identification via the use of a Naive Bayes classifier.
In proceedings of BIBE, 2008.

A. Oulas, A. Boutla, K. Gkirtzou, M. Reczko, K. Kalantidis and P. Poirazi.
Prediction of novel microRNA genes in cancer associated genomic regions a combined computational and experimental approach.
In press, Nucleic Acids Research.

K. Gkirtzou, P. Tsakalides and P. Poirazi.
MatureFind: a tool for identifying mature miRNAs in mammalians precursors.
Manuscript in preparation.