

# Efficient algorithms for the identification of miRNA motifs in DNA sequences

Nuno D. Mendes

Université Claude Bernard Lyon 1

Instituto Superior Técnico / Universidade Técnica de Lisboa

June 6th, 2011

**Supervisors:** Ana Teresa Correia de Freitas  
Marie-France Sagot



Finding pre-miRNAs

2011-06-06

Efficient algorithms for the identification of miRNA motifs in DNA sequences

Nuno D. Mendes

Université Claude Bernard Lyon 1

Instituto Superior Técnico / Universidade Técnica de Lisboa

June 6th, 2011

Supervisors: Ana Teresa Correia de Freitas

Marie-France Sagot



I would like to thank the members of the jury for having accepted being part of this committee.

I would also like to thank the people in the audience for being here today.

I will now present the results of my doctoral work, with a thesis titled 'Efficient algorithms for the identification of miRNA motifs in DNA sequences'.

This work was supervised by Ana Teresa Freitas from Instituto Superior Técnico and Marie-France Sagot from Université Claude Bernard Lyon 1.

# Part I

## Motivation

## Finding pre-miRNAs

- 2011-06-06

- ▶ The transcribed portion of the genome greatly exceeds the coding sequence
- ▶ Conservative estimates for the Human genome: (Cawthon, Estabrook, 2003)
  - Protein Coding Transcripts → 1.2% of the **euchromatic genome**
  - Transcribed Sequences → 40-70% of the **genome**
- ▶ Only a small portion of these non-coding transcripts (ncRNA) has been characterized
- ▶ Non-coding genomics ≠ Non-essential biology

Estimates for the human genome refer that protein coding transcripts make up for about 1.2% of the euchromatic genome, whereas 60 to 70% of the genome is transcribed

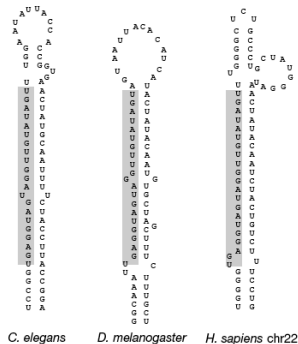
The topic of this thesis is an attempt at trying to better characterise a small portion of these non-coding transcripts: the microRNAs.

## Part II

## MicroRNA Biology

## Discovery of MicroRNAs

- ▶ Founding members of the **miRNA** class of non-coding RNAs were *lin-4* and *let-7* of *C. elegans*
  - ▶ *lin-4* and *let-7* are both 21-nt RNA sequences complementary to sites in the 3' UTR of genes they negatively regulate
  - ▶ Both were in a genomic context that could form extended stem-loops
  - ▶ *let-7* can be found in RNA samples from several different animal species



Pasquinelli et al., *Nature* 2000

## Finding pre-miRNAs

2011-06-06

└─ Discovery of MicroRNAs

MiRNAs were first identified in the nematode *C. elegans*. They were short RNA sequences with about 21-nt, complementary to sites in the 3' UTR region of coding genes which they were shown to regulate negatively.

In both cases, these short sequences were inserted in a genomic context where they formed a stem-loop like the ones shown in the figure.

Additionally, one of these founding miRNAs – *let-7* – was shown to be conserved across several animal species, hinting at the possibility that these regulators were ancient and phylogenetically extensive.

- Founding members of the **miRNA** class of non-coding RNAs were *lin-4* and *let-7* of *C. elegans*
  - lin-4* and *let-7* are both 21-nt RNA sequences complementary to sites in the 3' UTR of genes they negatively regulate
  - Both were in a genomic context that could form extended stem-loops
  - let-7* can be found in RNA samples from several different animal species





# MicroRNA Gene Finding

## Difficulties

### Finding pre-miRNAs

2011-06-06

## MicroRNA Gene Finding

MicroRNA Gene Finding

Difficulties

Limitations of experimental methods

- Some miRNAs have very low expression rates, or are only expressed in particular tissues/conditions
- Deep sequencing methods require extensive computational analyses

Limitations of computational approaches

- Mature miRNAs are too small to use conventional sequence analysis tools
- Non-coding genes have no obvious statistical properties as those explored in classical gene finding tools

## Limitations of experimental methods

- ▶ Some miRNAs have very low expression rates, or are only expressed in particular tissues/conditions
- ▶ Deep sequencing methods require extensive computational analyses

## Limitations of computational approaches

- ▶ Mature miRNAs are too small to use conventional sequence analysis tools
- ▶ Non-coding genes have no obvious statistical properties as those explored in classical gene finding tools

The identification of miRNA genes is met with several difficulties, illustrated by the fact that they took so long to be discovered in the first place.

First, some miRNAs have very low expression rates or are only expressed in particular tissues or particular physiological conditions. Although recent **deep-sequencing** techniques partly address this problem, they require extensive computational analyses and are subject to strong sequence biases that are not entirely characterised.

# MicroRNA Gene Finding

## Difficulties

## Finding pre-miRNAs

2011-06-06

## MicroRNA Gene Finding

MicroRNA Gene Finding

Difficulties

Limitations of experimental methods

- Some miRNAs have very low expression rates, or are only expressed in particular tissues/conditions
- Deep sequencing methods require extensive computational analyses

Limitations of computational approaches

- Mature miRNAs are too small to use conventional sequence analysis tools
- Non-coding genes have no obvious statistical properties as those explored in classical gene finding tools

## Limitations of experimental methods

- ▶ Some miRNAs have very low expression rates, or are only expressed in particular tissues/conditions
- ▶ Deep sequencing methods require extensive computational analyses

## Limitations of computational approaches

- ▶ Mature miRNAs are too small to use conventional sequence analysis tools
- ▶ Non-coding genes have no obvious statistical properties as those explored in classical gene finding tools

Standard gene finding approaches are not very useful, since the mature sequence of miRNAs are too short for conventional sequence analysis and non-coding genes in general do not exhibit the statistical properties of coding sequences that are at the heart of classical gene finding tools



# MicroRNA Gene Finding

## Early Approaches

## Finding pre-miRNAs

2011-06-06

## MicroRNA Gene Finding

MicroRNA Gene Finding

Early Approaches

Goal

Find **hairpin structures** resembling those observed in *lin-4* and *let-7* of *C. elegans* in **intergenic regions** that are **conserved** across two close species (e.g. *C. elegans* and *C. briggsae*)

The first computational approaches soon showed that it was possible to find a huge quantity of **conserved stem-loops** and more stringent criteria had to be used to sieve out miRNA candidates

## Goal

Find **hairpin structures** resembling those observed in *lin-4* and *let-7* of *C. elegans* in **intergenic regions** that are **conserved** across two close species (e.g. *C. elegans* and *C. briggsae*)

The first computational approaches soon showed that it was possible to find a huge quantity of **conserved stem-loops** and more stringent criteria had to be used to sieve out miRNA candidates

Early approaches to miRNA gene finding tried to identify stem-loops that had similar features to those initially identified in *C. elegans*.

They would have to be located in intergenic regions and show conservation across two close species.

The major lesson from these attempts was that genomes exhibited a vast quantity of conserved stem-loops and additional criteria had to be found in order to identify miRNA candidates.

# MicroRNA Gene Finding

## Current Approaches

2011-06-06

Finding pre-miRNAs

MicroRNA Gene Finding

MicroRNA Gene Finding  
Current Approaches

- Filter-based methods
- Machine learning approaches
- Target-centered approaches
- Homology-based searches

We have identified 4 major categories of modern miRNA gene finding tools, that we will now briefly describe

- 1 Filter-based methods
- 2 Machine learning approaches
- 3 Target-centered approaches
- 4 Homology-based searches

# MicroRNA Gene Finding

## Filter-based methods

- ▶ Identify an initial set of candidates using a given criterion (e.g. conserved or stable genomic stem-loops)
- ▶ Apply structural filters (e.g. HMM models, log-odds scoring schemes)
- ▶ Apply conservation filters (e.g. divergence patterns)

## Limitations

- ▶ Current methods are **biased** towards **highly conserved** pre-miRNAs with **structural details** similar to known miRNAs
- ▶ There is growing evidence for **non-conserved** miRNAs which are either clade or species-specific

## Finding pre-miRNAs

2011-06-06

### MicroRNA Gene Finding

#### MicroRNA Gene Finding

##### Filter-based methods

- ▶ Identify an initial set of candidates using a given criterion (e.g. conserved or stable genomic stem-loops)
- ▶ Apply structural filters (e.g. HMM models, log-odds scoring schemes)
- ▶ Apply conservation filters (e.g. divergence patterns)

##### Limitations

- ▶ Current methods are biased towards highly conserved pre-miRNAs with structural details similar to known miRNAs
- ▶ There is growing evidence for non-conserved miRNAs which are either clade or species-specific

Filter-based methods begin by identifying an initial set of candidates and then apply structural filters, to select only candidates with particular structural features and conservation filters to detect sequence/structure conservation or a particular **divergence pattern**.

# MicroRNA Gene Finding

## Filter-based methods

- ▶ Identify an initial set of candidates using a given criterion (e.g. conserved or stable genomic stem-loops)
- ▶ Apply structural filters (e.g. HMM models, log-odds scoring schemes)
- ▶ Apply conservation filters (e.g. divergence patterns)

## Limitations

- ▶ Current methods are **biased** towards **highly conserved** pre-miRNAs with **structural details** similar to known miRNAs
- ▶ There is growing evidence for **non-conserved** miRNAs which are either clade or species-specific

## Finding pre-miRNAs

2011-06-06

## MicroRNA Gene Finding

MicroRNA Gene Finding

Filter-based methods

- ▶ Identify an initial set of candidates using a given criterion (e.g. conserved or stable genomic stem-loops)
- ▶ Apply structural filters (e.g. HMM models, log-odds scoring schemes)
- ▶ Apply conservation filters (e.g. divergence patterns)

**Limitations**

- ▶ Current methods are **biased** towards **highly conserved** pre-miRNAs with **structural details** similar to known miRNAs
- ▶ There is growing evidence for **non-conserved** miRNAs which are either clade or species-specific

These approaches are generally biased towards highly-conserved precursors and the structural details used to filter the candidates rely on those observed in currently known miRNAs

However, there is evidence of non-conserved miRNAs and the structural requirements for miRNA recognition may not be easily generalised from current known precursors

# MicroRNA Gene Finding

## Machine learning approaches

- ▶ Try to learn the defining characteristics of known miRNAs
- ▶ Use a sequence/structure features and global properties like entropy, MFE, and conservation patterns

### Training Sets

- Positive** All known microRNA precursors
- Negative** tRNAs, rRNAs, and other stem-loops randomly recovered from the genome

### Limitations

- ▶ Positive set is biased towards **highly-expressed, highly-conserved** miRNAs identified by previous experimental and computational methods
- ▶ It is uncertain how many structures in the negative set could be processed by the miRNA maturation pathway

## Finding pre-miRNAs

2011-06-06

### MicroRNA Gene Finding

#### MicroRNA Gene Finding

##### Machine learning approaches

- ▶ Try to learn the defining characteristics of known miRNAs
- ▶ Use a sequence/structure features and global properties like entropy, MFE, and conservation patterns

##### Training sets

- Positive** All known microRNA precursors
- Negative** tRNAs, rRNAs, and other stem-loops randomly recovered from the genome

##### Limitations

- ▶ Positive set is biased towards highly-expressed, highly-conserved miRNAs identified by previous experimental and computational methods
- ▶ It is uncertain how many structures in the negative set could be processed by the miRNA maturation pathway

Machine learning approaches rely on sequence/structure features or global properties of known miRNAs versus other genomic stem-loops

# MicroRNA Gene Finding

## Machine learning approaches

- ▶ Try to learn the defining characteristics of known miRNAs
- ▶ Use a sequence/structure features and global properties like entropy, MFE, and conservation patterns

## Training Sets

- |                 |   |
|-----------------|---|
| <b>Positive</b> | All known microRNA precursors   |
| <b>Negative</b> | tRNAs, rRNAs, and other stem-loops randomly recovered from the genome |

## Limitations

- ▶ Positive set is biased towards **highly-expressed, highly-conserved** miRNAs identified by previous experimental and computational methods
- ▶ It is uncertain how many structures in the negative set could be processed by the miRNA maturation pathway

## Finding pre-miRNAs

2011-06-06

### MicroRNA Gene Finding

MicroRNA Gene Finding

Machine learning approaches

- ▶ Try to learn the defining characteristics of known miRNAs
- ▶ Use a sequence/structure features and global properties like entropy, MFE, and conservation patterns

**Training Sets**

<b>Positive</b>	All known microRNA precursors
<b>Negative</b>	tRNAs, rRNAs, and other stem-loops randomly recovered from the genome

**Limitations**

- ▶ Positive set is biased towards highly-expressed, highly-conserved miRNAs identified by previous experimental and computational methods
- ▶ It is uncertain how many structures in the negative set could be processed by the miRNA maturation pathway

The training set uses known precursors as positive examples and other genomic stem-loops (either from other non-coding transcripts or randomly recovered hairpins) as negative examples.

# MicroRNA Gene Finding

## Machine learning approaches

- ▶ Try to learn the defining characteristics of known miRNAs
- ▶ Use a sequence/structure features and global properties like entropy, MFE, and conservation patterns

## Training Sets

- Positive** All known microRNA precursors
- Negative** tRNAs, rRNAs, and other stem-loops randomly recovered from the genome

## Limitations

- ▶ Positive set is biased towards **highly-expressed, highly-conserved** miRNAs identified by previous experimental and computational methods
- ▶ It is uncertain how many structures in the negative set could be processed by the miRNA maturation pathway

## Finding pre-miRNAs

2011-06-06

### MicroRNA Gene Finding

MicroRNA Gene Finding

Machine learning approaches

- ▶ Try to learn the defining characteristics of known miRNAs
- ▶ Use a sequence/structure features and global properties like entropy, MFE, and conservation patterns

**Training Sets**

- Positive** All known microRNA precursors
- Negative** tRNAs, rRNAs, and other stem-loops randomly recovered from the genome

**Limitations**

- ▶ Positive set is biased towards **highly-expressed, highly-conserved** miRNAs identified by previous experimental and computational methods
- ▶ It is uncertain how many structures in the negative set could be processed by the miRNA maturation pathway

The positive set is biased towards highly-expressed and highly-conserved miRNAs which may not be representative of the structures that can, in principle, be recognised as precursors, and the negative set is not guaranteed to include only structures which would not enter the miRNA maturation pathway.

# MicroRNA Gene Finding

## Target-centered approaches

- ▶ Potential target sites are identified by looking for **highly conserved motifs** in potential target regions
- ▶ New miRNAs are sought by identifying **conserved stem-loops** with **complementary sequences**

### Finding pre-miRNAs

2011-06-06

└─MicroRNA Gene Finding

Target-centered approaches rely on the identification of conserved target sites in genes potentially targetted by miRNAs, which are then used to identify conserved genomic hairpins with complementary sequences.



# MicroRNA Gene Finding

## Target-centered approaches

- ▶ Potential target sites are identified by looking for **highly conserved motifs** in potential target regions
- ▶ New miRNAs are sought by identifying **conserved stem-loops** with **complementary sequences**

### Limitations

- ▶ They have the advantage of making few assumptions about precursor structure, **but**
- ▶ Depend on the identification of highly conserved target sites

## Finding pre-miRNAs

2011-06-06

### MicroRNA Gene Finding

MicroRNA Gene Finding

Target-centered approaches

- ▶ Potential target sites are identified by looking for **highly conserved motifs** in potential target regions
- ▶ New miRNAs are sought by identifying **conserved stem-loops** with **complementary sequences**

**Limitations**

- ▶ They have the advantage of making few assumptions about precursor structure, **but**
- ▶ Depend on the identification of highly conserved target sites

These approaches have the advantage of making few assumptions about the structure of potential precursors (except that they ought to be conserved), but they are critically dependent on the identification of conserved target sites, which is a challenge on its own.

## Homology-based searches

2011-06-06

└─MicroRNA Gene Finding

Finally, homology-based searches purport to identify homologs of known precursors by considering a mixture of sequence/structure conservation measures.

# MicroRNA Gene Finding

## Homology-based searches

### Finding pre-miRNAs

2011-06-06

└─MicroRNA Gene Finding

MicroRNA Gene Finding

### Homology-based searches

- ▶ Use alignment-based methods to find homologs to previously known miRNAs
- ▶ Consider a mixture of sequence/structure conservation measures

### Limitations

- Are limited to identifying relatively close homologs
- Cannot find new families of miRNAs

- ▶ Use alignment-based methods to find homologs to previously known miRNAs
- ▶ Consider a mixture of sequence/structure conservation measures

## Limitations

- ▶ Are limited to identifying relatively close homologs
- ▶ Cannot find new families of miRNAs

These methods are limited to identifying relatively close homologs and cannot hope to find new families of miRNAs.

## Part IV

### CRAVELA Framework

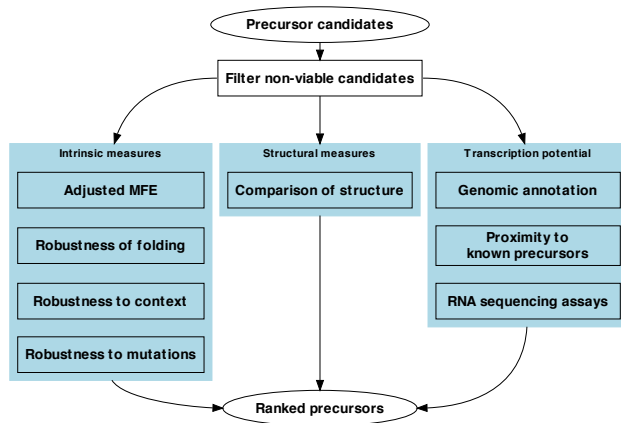
**Publications:** N D Mendes, A T Freitas, A T Vasconcelos, M-F Sagot.  
*Combination of measures distinguishes pre-miRNAs from other stem-loops in the genome of the newly sequenced Anopheles darlingi*, BMC Genomics, 2010  
N D Mendes, S Heyne, A T Freitas, M-F Sagot, R Backofen  
*Navigating the unexplored seascape of pre-miRNA candidates in single-genome approaches*, In preparation



We now present our own proposal to address the problem of identifying miRNA genes. The material we will now present is supported by two papers, one is already published in BMC Genomics and the other is in preparation.

# CRAVELA Framework

## Classification of candidates

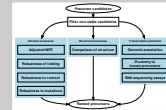


## Finding pre-miRNAs

2011-06-06

CRAVELA Framework

CRAVELA Framework  
Classification of candidates



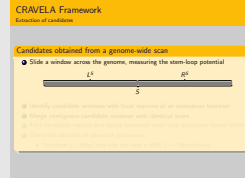
The CRAVELA framework is a tool to enumerate and score miRNA precursor candidates from a **single-genome**.

It begins by identifying a large set of potential precursors which are then subjected to three types of evaluation.

First, evaluation w.r.t. intrinsic measures of **stability and robustness**

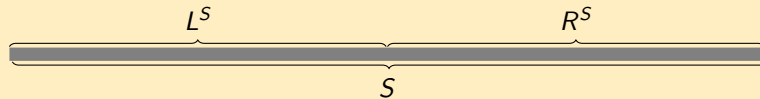
Then a structural analysis which purports to identify **structural classes** within the candidates, particularly candidates which are **structurally similar** to known pre-miRNAs.

And, finally, an evaluation of the transcriptional potential of these candidates by profiting from whatever data are available for the genome under study.



## Candidates obtained from a genome-wide scan

- 1 Slide a window across the genome, measuring the stem-loop potential



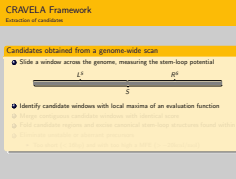
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
  - ▶ Too short ( $< 16\text{bp}$ ) and with too high a MFE ( $> -20\text{kcal/mol}$ )

The first step in the CRAVELA processing pipeline is the extraction of the initial set of candidates

To accomplish this, we slide a window across the genome, which is large enough to accommodate even large precursors sequences.

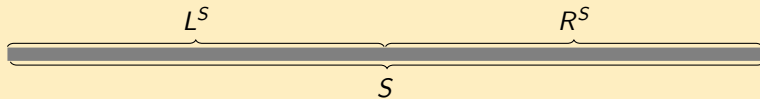
Instead of folding the sequence within the window at each step, we simply evaluate the hybridisation potential of both halves of the window using an evaluation function based on complementary alignment scores.  
(This is equivalent to folding the sequence with a very simple energy model and with an additional restriction concerning which segments can be paired with each other)

This procedure is much less computationally expensive than RNA folding



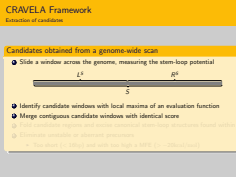
## Candidates obtained from a genome-wide scan

- 1 Slide a window across the genome, measuring the stem-loop potential



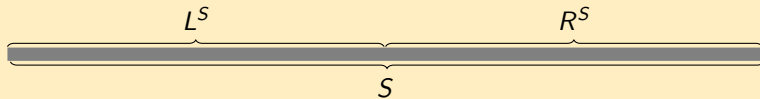
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
  - ▶ Too short ( $< 16\text{bp}$ ) and with too high a MFE ( $> -20\text{kcal/mol}$ )

As we slide across the genome, we only retain **candidate windows** which correspond to local maxima of the evaluation function



## Candidates obtained from a genome-wide scan

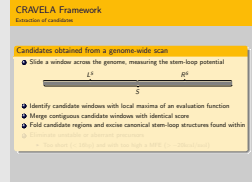
- 1 Slide a window across the genome, measuring the stem-loop potential



- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
  - ▶ Too short ( $< 16\text{bp}$ ) and with too high a MFE ( $> -20\text{kcal/mol}$ )

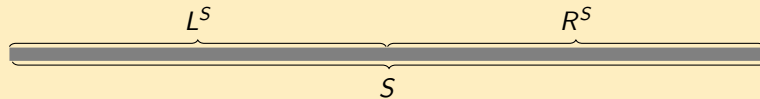
Contiguous candidate windows with identical scores are merged into candidate regions





## Candidates obtained from a genome-wide scan

- 1 Slide a window across the genome, measuring the stem-loop potential



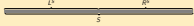
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
  - ▶ Too short ( $< 16\text{bp}$ ) and with too high a MFE ( $> -20\text{kcal/mol}$ )

Candidate regions are folded using RNAfold and the largest stem-loop contained therein is identified

CRAVELA Framework  
Extraction of candidates

Candidates obtained from a genome-wide scan

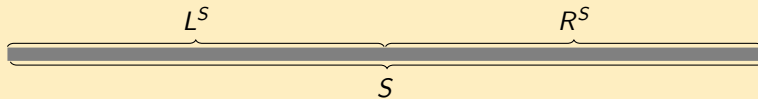
- 1 Slide a window across the genome, measuring the stem-loop potential



- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
  - Too short ( $< 16\text{bp}$ ) and with too high a MFE ( $> -20\text{kcal/mol}$ )

## Candidates obtained from a genome-wide scan

- 1 Slide a window across the genome, measuring the stem-loop potential



- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
  - ▶ Too short ( $< 16\text{bp}$ ) and with too high a MFE ( $> -20\text{kcal/mol}$ )

Short or unstable stem-loops are discarded

# CRAVELA Framework

## Extraction of candidates

## Finding pre-miRNAs

2011-06-06

CRAVELA Framework

CRAVELA Framework	
Extraction of candidates	
Number of candidates	
<i>Drosophila melanogaster</i>	1 316 305
<i>Anopheles gambiae</i>	2 245 014
<i>Anopheles darlingi</i>	1 748 153
Number of known precursors	
<i>Drosophila melanogaster</i>	157
<i>Anopheles gambiae</i>	67
<i>Anopheles darlingi</i>	44 (identified by homology to <i>A. gambiae</i> )

## Number of candidates

<i>Drosophila melanogaster</i>	1 316 305
<i>Anopheles gambiae</i>	2 245 014
<i>Anopheles darlingi</i>	1 748 153

For the three metazoan genomes we studied, we obtained candidates in the order of a million compared to about **a hundred or less** documented miRNAs in these genomes.

In the case of *A. darlingi* these precursors were identified by homology and they are practically identical to precursors documented in *A. gambiae*

## Number of known precursors

<i>Drosophila melanogaster</i>	157
<i>Anopheles gambiae</i>	67
<i>Anopheles darlingi</i>	44 (identified by homology to <i>A. gambiae</i> )

## Assessing the performance of intrinsic measures

- 2011-06-06

└─ CRAVELA Framework

## CRAVELA Framework

- ▶ Undersampling procedure in order to produce 1000 samples with an identical number of known precursors (constant) and non-overlapping candidates
- ▶ ROC curve for each sample reflecting the trade-off between specificity/sensitivity with respect to cutoff level
- ▶ Optimal cutoff determined using the Youden index (i.e. the cutoff which maximises Specificity + Sensitivity)

The trade-off between specificity/sensitivity for each measure on each sample is determined by a ROC curve.

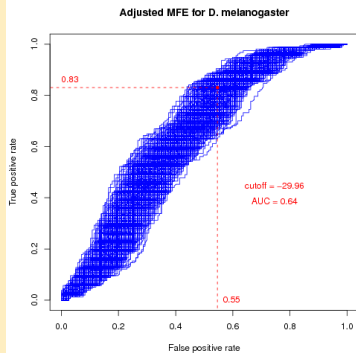
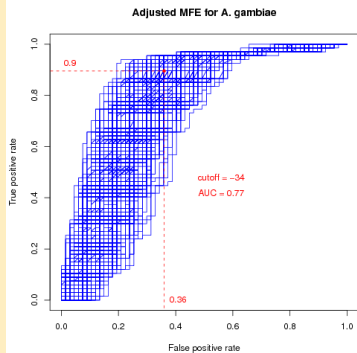
The optimal cutoff in each sample is calculated using the Youden index, which corresponds to the cutoff value that maximises the sum of specificity and sensitivity.

# CRAVELA Framework

## Intrinsic Measures

### Adjusted MFE (Zhang et al., 2006)

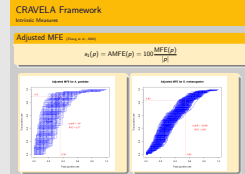
$$s_1(p) = \text{AMFE}(p) = 100 \frac{\text{MFE}(p)}{|p|}$$



## Finding pre-miRNAs

2011-06-06

CRAVELA Framework



The adjusted minimum free energy is an evaluation measure already shown to distinguish stem-loops originating from miRNAs from stem-loops coming from other ncRNAs.

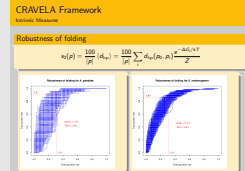
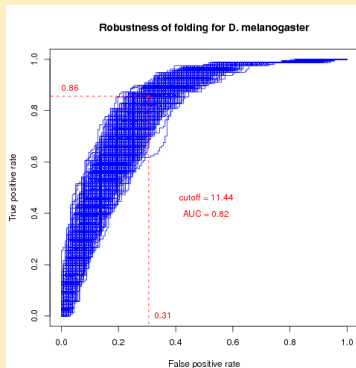
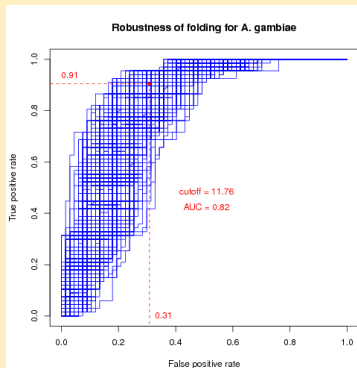
It is simply the normalisation of the MFE by the length of the candidate.

In the figure we can see the ROC curves for the datasets of *A. gambiae* and *D. melanogaster*, showing the average true/false positive rates over all samples for each respective optimal cutoff.

The performance of this measure is somewhat poor in the *D. melanogaster* dataset, possibly due to the inclusion of heterochromatic sequences and the fact that this measure does not compensate for GC content.

### Robustness of folding

$$s_2(p) = \frac{100}{|p|} \langle d_{bp} \rangle = \frac{100}{|p|} \sum_i d_{bp}(p_0, p_i) \frac{e^{-\Delta G_i/kT}}{Z}$$



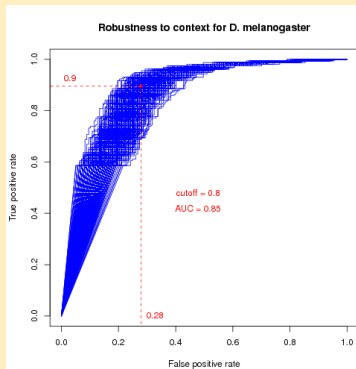
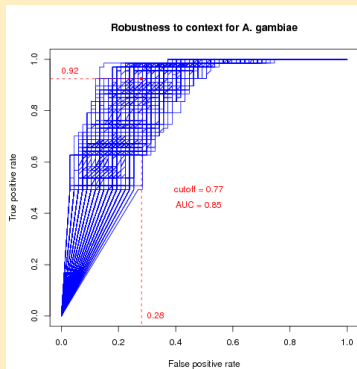
The robustness of folding is a measure that purports to assess whether the secondary structure of stem-loops varies much across the suboptimal structures of its thermodynamic ensemble. Should the ensemble include vary diverse structures, it is unlikely that the MFE structure is stable enough to guarantee recognition by the enzymes involved in miRNA biogenesis.

This measure is calculated as the average of the basepair distance between the MFE structure and each suboptimal structure, weighted by their probability

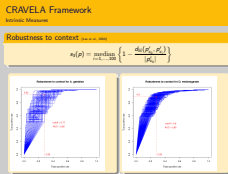
As is shown in the figures, the measure has comparable performances in both datasets, with an average AUC of 0.82

### Robustness to context (Lee et al., 2008)

$$s_3(p) = \text{median}_{i=1, \dots, 100} \left\{ 1 - \frac{d_H(p'_{c_0}, p'_{c_i})}{|p'_{c_0}|} \right\}$$



CRAVELA Framework



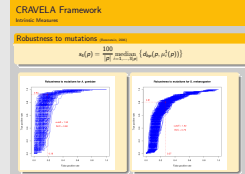
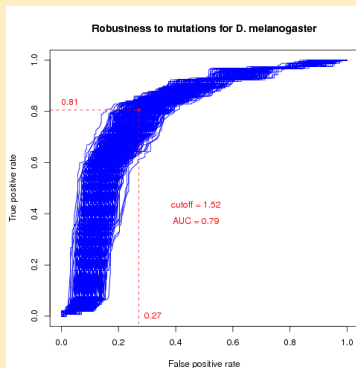
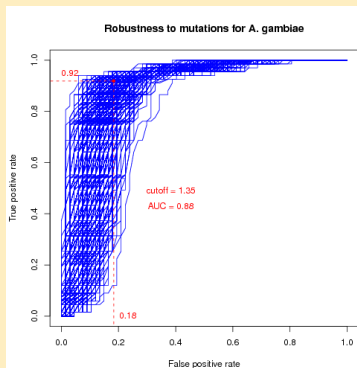
The robustness to context is a measure of how resilient a stem-loop is to variations in its genomic context.

We calculate this measure as the median proportion of the secondary structure of the candidate, folded in its original context, that is preserved across 100 random genomic contexts that exhibit the same dinucleotide frequencies than the original context.

Once again, performances are similar in both datasets, with an AUC of 0.85.

### Robustness to mutations (Borenstein, 2006)

$$s_4(p) = \frac{100}{|p|} \text{median}_{i=1, \dots, 3|p|} \{d_{bp}(p, \mu_i^1(p))\}$$

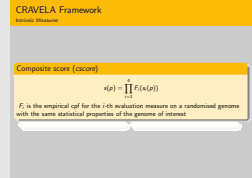


The robustness to mutations measures the impact of point mutations in the structure of the precursor candidates.

It is calculated as the median basepair distance between the original structure and the structure resulting from all possible point mutations in the candidate sequence.

The figures show a better performance for the *A. gambiae* dataset. This might be explained by the fact that candidates arising from the heterochromatic sequences in *D. melanogaster*, which tend to be very repetitive, are also very resilient to point mutations.





## Composite score (*cscore*)

$$s(p) = \prod_{i=1}^4 F_i(s_i(p))$$

$F_i$  is the empirical cpf for the  $i$ -th evaluation measure on a randomised genome with the same statistical properties of the genome of interest

The procedure used to combine all the intrinsic measures involves two steps.

The first step consists in generating a random genome which preserves the dinucleotide frequencies of the original genome. Candidate structures are extracted from this artificial genome and each of the four measures are calculated for these hairpins. The distribution of the measures in this artificial genome is used as a background distribution against which the scores of our candidates are compared.

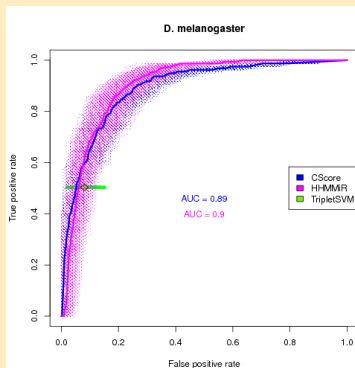
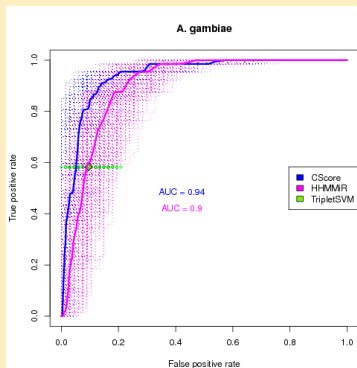
We recast the score of each measure as the value of the cumulative probability function of the empirical distribution of each measure in the artificial genome *and* we determine the combined score as the product of the scores of the four measures.

# CRAVELA Framework

## Intrinsic Measures

### Composite score (*cscore*)

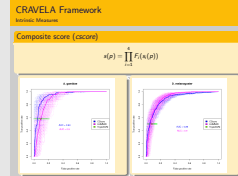
$$s(p) = \prod_{i=1}^4 F_i(s_i(p))$$



## Finding pre-miRNAs

2011-06-06

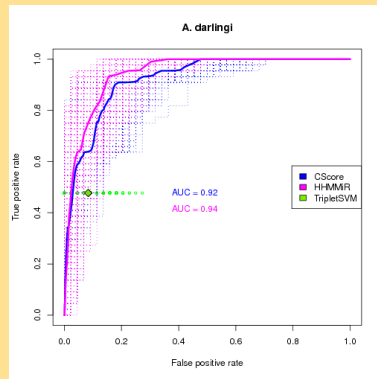
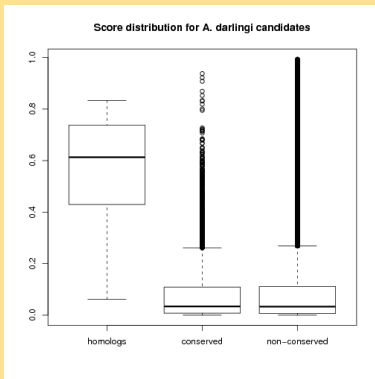
└ CRAVELA Framework



In the figures, we show the ROC curves for our combined score along the the ROC curves for another single-genome method, which is a HMM-based approach trained on human pre-miRNAs, and additionally we show the performance of an SVM method TripletSVM, also trained on human precursors, across our samples.

Our **unsupervised** approach outperforms HHMMIR in the *A. gambiae* dataset and has comparable performance in the *D. melanogaster* dataset. In both cases it is superior to the average performance of TripletSVM.

## Exploration of candidates in *A. darlingi*



We can see that our approach also behaves well for the *A. darlingi* dataset and in this case we can make an additional observation.

The graph on the left shows the distribution of our combined score in three classes of hairpins. Homologs, conserved stem-loops across the two *Anopheles* species, and the non-conserved or low-conservation stem-loops.

We see that conservation has practically no impact on the distribution of the combined score, which means that conserved stem-loops are not necessarily more or less robust. And that conservation alone is a not very good way to assess whether a stem-loop is a good miRNA precursor candidate.

2011-06-06

Finding pre-miRNAs

CRAVELA Framework

CRAVELA Framework			
Intrinsic measures			
Number of candidates above cut-off			
<i>Drosophila melanogaster</i>	240 751		
<i>Anopheles gambiae</i>	328 829		
<i>Anopheles darlingi</i>	305 681		
		Cut-off	Sensitivity
			Specificity
<i>Drosophila melanogaster</i>		0.30	0.83
<i>Anopheles gambiae</i>		0.41	0.90
<i>Anopheles darlingi</i>		0.32	0.89

## Number of candidates above cut-off

<i>Drosophila melanogaster</i>	240 751
<i>Anopheles gambiae</i>	328 829
<i>Anopheles darlingi</i>	305 681

The combined score elicits the reduction of the number of candidate by one order of magnitude to the order of a hundred thousand

and with generally high sensitivity and good specificity.

	Cut-off	Sensitivity	Specificity
<i>Drosophila melanogaster</i>	0.30	0.83	0.80
<i>Anopheles gambiae</i>	0.41	0.90	0.88
<i>Anopheles darlingi</i>	0.32	0.89	0.84

# CRAVELA Framework

Intrinsic measures

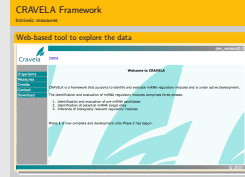
## Web-based tool to explore the data



Finding pre-miRNAs

2011-06-06

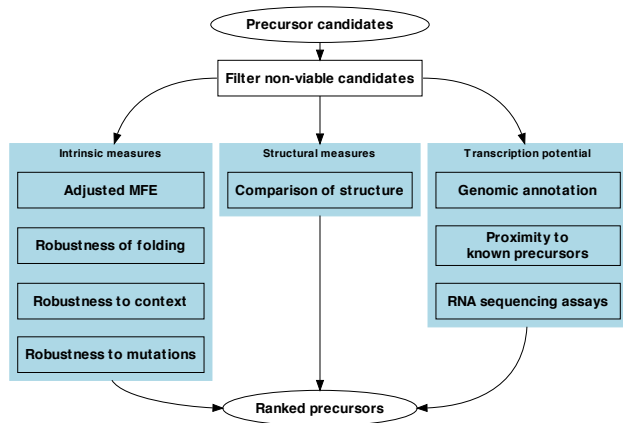
CRAVELA Framework



The results of the intrinsic measures analysis are partially made available in a web-based tool, which is under development, and will soon be extended to incorporate the results of the other evaluation steps as well.

# CRAVELA Framework

## Classification of candidates



## Finding pre-miRNAs

2011-06-06

CRAVELA Framework

CRAVELA Framework  
Classification of candidates



Now we describe the next evaluation step: structural analysis

### Goal

Identify which structures are most likely to be recognised and processed as pre-miRNAs

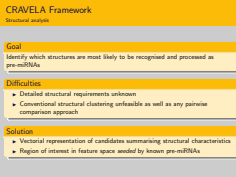
### Difficulties

- ▶ Detailed structural requirements unknown
- ▶ Conventional structural clustering unfeasible as well as any pairwise comparison approach

### Solution

- ▶ Vectorial representation of candidates summarising structural characteristics
- ▶ Region of interest in feature space *seeded* by known pre-miRNAs

CRAVELA Framework



The goal of our structural analysis is to try to identify what structures are likely to be recognised as pre-miRNAs.

However, we ignore the precise structural requirements.

We could try to cluster our candidates using conventional structural clustering tools in search for salient structural classes, but the number of structures (several hundred thousand) make it computationally unfeasible.

The approach we propose is to find a way to represent the sequence/structure features of our candidates and to try to identify the region of the feature space most likely to harbour candidates with the adequate structural characteristics.

This region of interest is to be seeded by known precursors.

### Vectorial Representation

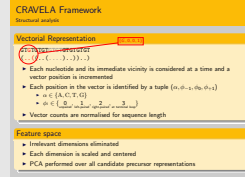
(G, 0, 0, 1)

GTGTCGTGTCGTGTCGTGTCGT  
 (...((...(...))...))...)

- ▶ Each nucleotide and its immediate vicinity is considered at a time and a vector position is incremented
- ▶ Each position in the vector is identified by a tuple  $(\alpha, \phi_{-1}, \phi_0, \phi_{+1})$ 
  - ▶  $\alpha \in \{A, C, T, G\}$
  - ▶  $\phi_i \in \{ \underset{\text{unpaired}}{0}, \underset{\text{left-paired}}{1}, \underset{\text{right-paired}}{2}, \underset{\text{at terminal loop}}{3} \}$
- ▶ Vector counts are normalised for sequence length

### Feature space

- ▶ Irrelevant dimensions eliminated
- ▶ Each dimension is scaled and centered
- ▶ PCA performed over all candidate precursor representations



Our representation of the sequence/structure features of each candidate involves considering each position in the stem-loop and to keep a vector of counts for each time a particular nucleotide appears in a given structural context.

The positions in the vector are given by a tuple that identifies which nucleotide is present, as well as the pairing state of the previous, the actual and the next position in the stem-loop. Pairing states can be **unpaired**, **left paired**, **right paired** or **unpaired in the terminal loop**.

The counts are divided by the length of the precursor.

Vector positions with zeroes are removed and each vector position is scaled and centered around the mean before a PCA analysis is performed.

The coordinates of each candidate in the feature space are given by the principal components.

This allows us to have a space of linearly independent dimensions where we can compute meaningful Euclidian distances



CRAVELA Framework  
Structural analysis

Feature space relative positions reflect structural similarity

- Several **small** samples taken from datasets (100 samples with 1000 stem-loops)
- Structural clusters in these samples found using a conventional structural clustering approach (LocARNA)
- Candidate positions in the feature space are much closer to their cluster centroid than what would be expected by chance
- Known precursors in the samples are also closer to each other than expected by chance

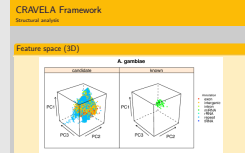
## Feature space relative positions reflect structural similarity

- ▶ Several **small** samples taken from datasets (100 samples with 1000 stem-loops)
- ▶ Structural clusters in these samples found using a conventional structural clustering approach (LocARNA)
- ▶ Candidate positions in the feature space are much closer to their cluster centroid than what would be expected by chance
- ▶ Known precursors in the samples are also closer to each other than expected by chance

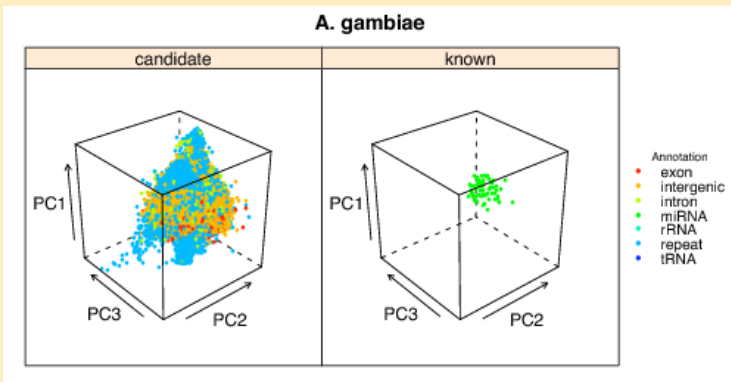
To assess whether our feature space captures actual structural similarity we take 100 samples of 1000 stem-loops and we cluster them using a conventional structural clustering approach.

Candidates that belong to the same **structural cluster** ought to be closer in the feature space, We determine this by calculating the **centroid** in the feature space of each structural cluster and verify whether each cluster member is closer to the centroid of its own cluster than to the centroid of any other cluster. Using a **randomisation** procedure, we verify that this is the case, and that candidates tend to be much closer to their centroid than what would be expected by chance.

Additionally, we also verify that known precursors are much closer to each other than expected by chance.



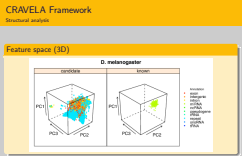
## Feature space (3D)



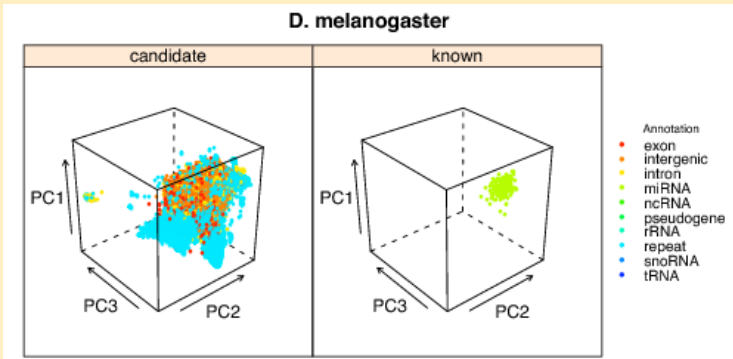
If we restrict the feature space to the first three principal components for the purposes of visualisation we can see by inspection that in *A. gambiae* known precursors are indeed restricted to a limited portion of the feature space.

2011-06-06

Finding pre-miRNAs



## Feature space (3D)



The same is true for the D. melanogaster dataset.

- ▶ Known precursors are close to each other **but** within dense region
- ▶ Region of interest identified at the expense of known precursors

### Acceptance region (MinDist)

A precursor candidate  $c$  is in the acceptance region iff

$$\min_{p \in \mathcal{P}} d(c, p) < r, \quad r > 0$$

where  $\mathcal{P}$  is the set of known precursors, and  $d$  is the distance in the multidimensional feature space.

CRAVELA Framework

CRAVELA Framework  
Structural analysis

- ▶ Known precursors are close to each other **but** within dense region
- ▶ Region of interest identified at the expense of known precursors

**Acceptance region (MinDist)**  
A precursor candidate  $c$  is in the acceptance region iff

$$\min_{p \in \mathcal{P}} d(c, p) < r, \quad r > 0$$

where  $\mathcal{P}$  is the set of known precursors, and  $d$  is the distance in the multidimensional feature space.

Known precursors occupy a limited portion of the feature space but it is, however, a dense region, which means that many stem-loops seem to exhibit similar sequence/structure feature to those of known miRNAs.

This region – made up of the intersection of hyperspheres – is identified at the expense of the documented precursors using a procedure we call MinDist. The size of the region is controlled by the parameter  $r$ , which refers to the maximum distance allowed between a candidate and the closest precursor.

# CRAVELA Framework

## Assessing the performance of MinDist

- Performance of MinDist compared to that of TripletSVM
- Four groups of 1000 samples of stem-loops.
- Each group uses 5%, 10%, 20% and 50%, respectively, of the set of known precursors of each dataset for the training/seed set, and the remaining precursors and used in the test set.
- In both the training/seed sets and test sets, an identical number of examples (presumed negative) are samples from the set of candidates

## Finding pre-miRNAs

2011-06-06

CRAVELA Framework

CRAVELA Framework  
Assessing the performance of MinDist

- Performance of MinDist compared to that of TripletSVM
- Four groups of 1000 samples of stem-loops
- Each group uses 5%, 10%, 20% and 50%, respectively, of the set of known precursors of each dataset for the training/seed set, and the remaining precursors and used in the test set.
- In both the training/seed sets and test sets, an identical number of examples (presumed negative) are samples from the set of candidates

We assess the performance of our procedure and we compare it to the performance of TripletSVM

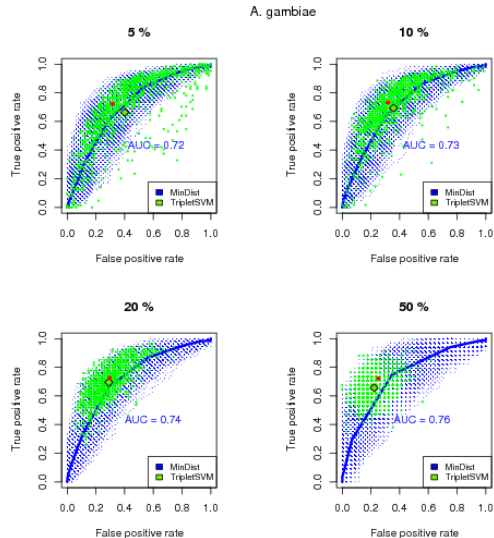
The performance of the procedure is assessed by considering several groups of 1000 samples of stem-loops.

Each group uses a varying percentage of the known precursors to make up the training set of TripletSVM or the seed set for MinDist, the remaining precursors are used in the test set.

In every set, there is an identical number of presumably negative examples taken from samples of the candidates.

# CRAVELA Framework

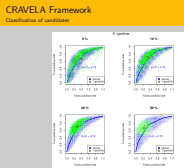
## Classification of candidates



## Finding pre-miRNAs

2011-06-06

CRAVELA Framework



These are the results for the *A. gambiae* dataset in sample groups containing 5%, 10%, 20% and 50% of known precursors for training or seeding. The

dashed blue lines are the ROC curves for each sample. The solid blue line is the average ROC curve. The red dot is the average performance of MinDist w.r.t. each sample's optimal cutoff.

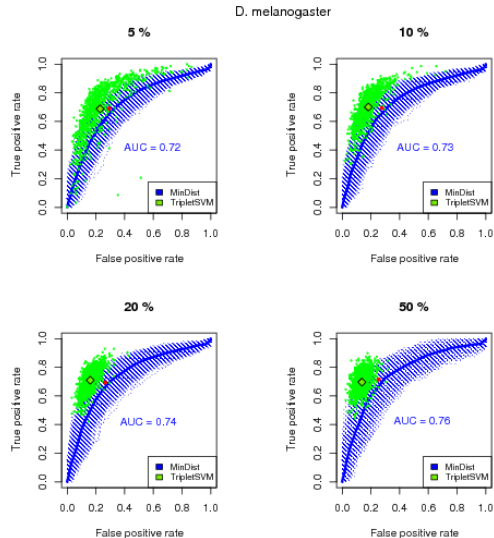
The green dots are the performance of TripletSVM in each sample, and the green diamond is its average performance.

In this dataset, MinDist outperforms TripletSVM in every sample group.

Additionally, we can see that TripletSVM has a very unstable performance in sample groups with a fewer number of examples, alternating between good and very poor performance.

# CRAVELA Framework

## Classification of candidates



## Finding pre-miRNAs

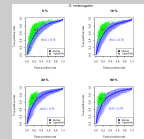
2011-06-06

CRAVELA Framework

The results are similar in the *D. melanogaster* dataset, except that TripletSVM tends to attain better specificity.

CRAVELA Framework

Classification of candidates



# CRAVELA Framework

## Classification of candidates

Finding pre-miRNAs

2011-06-06



CRAVELA Framework

Classification of candidates

% known	<i>A. gambiae</i>				<i>D. melanogaster</i>			
	MinDist	$F_1$	TripletSVM	$F_1$	MinDist	$F_1$	TripletSVM	$F_1$
5 %	0.72	0.68	0.71	0.64	0.69	0.71	0.70	0.72
10 %	0.73	0.68	0.71	0.67	0.69	0.72	0.70	0.74
20 %	0.73	0.68	0.71	0.70	0.69	0.74	0.71	0.76
50 %	0.72	0.75	0.73	0.72	0.71	0.75	0.72	0.76
80 %	0.74	0.80	0.76	0.70	0.72	0.77	0.74	0.76
90 %	0.78	0.83	0.80	0.73	0.73	0.81	0.76	0.76
95 %	0.75	0.89	0.81	0.71	0.75	0.83	0.78	0.76

Table: Sensitivity (TPR), Specificity (1 - FPR) and the  $F_1$  measure  $(2 \frac{TP}{TP+FP+TPR})$  of TripletSVM and MinDist computed as the average performance across all samples for training sets whose positive examples consist of a fraction of known pre-miRNAs in *A. gambiae* and *D. melanogaster*

% known	<i>A. gambiae</i>						<i>D. melanogaster</i>					
	MinDist			TripletSVM			MinDist			TripletSVM		
	Sens.	Spec.	$F_1$	Sens.	Spec.	$F_1$	Sens.	Spec.	$F_1$	Sens.	Spec.	$F_1$
5 %	0.72	0.68	0.71	0.66	0.60	0.64	0.69	0.71	0.70	0.69	0.77	0.72
10 %	0.73	0.68	0.71	0.69	0.64	0.67	0.69	0.72	0.70	0.70	0.82	0.74
20 %	0.73	0.68	0.71	0.69	0.71	0.70	0.69	0.74	0.71	0.71	0.84	0.76
50 %	0.72	0.75	0.73	0.66	0.78	0.72	0.71	0.75	0.72	0.70	0.86	0.76
80 %	0.74	0.80	0.76	0.65	0.80	0.70	0.72	0.77	0.74	0.69	0.88	0.76
90 %	0.78	0.83	0.80	0.65	0.88	0.73	0.73	0.81	0.76	0.68	0.88	0.76
95 %	0.75	0.89	0.81	0.66	0.81	0.71	0.75	0.83	0.78	0.68	0.88	0.76

**Table:** Sensitivity (TPR), Specificity (1 - FPR) and the  $F_1$  measure  $(2 \frac{TP}{TP+FP+TPR})$  of TripletSVM and MinDist computed as the average performance across all samples for training sets whose positive examples consist of a fraction of known pre-miRNAs in *A. gambiae* and *D. melanogaster*

This table summarises and extends our analysis to sample groups including up to 95% of known precursors.

Explain better

In the *A. gambiae* dataset MinDist is clearly superior to TripletSVM according to the  $F_1$  measure across all sample groups

For the *D. melanogaster* dataset the performance is comparable.



CRAVELA Framework

Structural analysis

Estimating optimal cut-off

- Each sample group average optimal cut-off and proportion of known precursors used follow a log-linear law until reaching about 90%
- Distance used to calculate acceptance region estimated by extrapolating the log-linear law to 100%

Number of candidates within acceptance region

<i>Drosophila melanogaster</i>	67 619
<i>Anopheles gambiae</i>	77 366

### Estimating optimal cut-off

- ▶ Each sample group average optimal cut-off and proportion of known precursors used follow a log-linear law until reaching about 90%
- ▶ Distance used to calculate acceptance region estimated by extrapolating the log-linear law to 100%

The optimal value for the distance parameter is extrapolated from the log-linear law followed by the relation between the optimal cut-off and the proportion of known precursors used in each sample group.

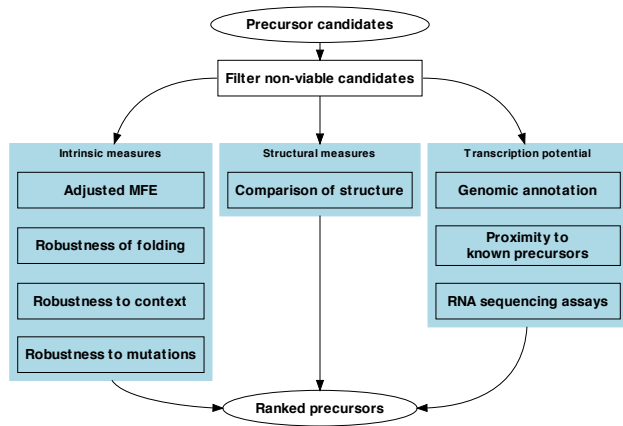
### Number of candidates within acceptance region

<i>Drosophila melanogaster</i>	67 619
<i>Anopheles gambiae</i>	77 366

The distance parameter calculated for each dataset defines an acceptance region which reduces the number of candidates by an order of magnitude to the order of tens of thousands.

# CRAVELA Framework

## Classification of candidates

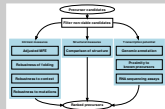


## Finding pre-miRNAs

2011-06-06

CRAVELA Framework

CRAVELA Framework  
Classification of candidates



The final analysis concerns the assessment of the transcriptional potential. We shall discuss in this presentation the analysis of annotation data and the search for miRNA genomic clusters.

### Annotation

- ▶ Annotation data provide information that can be used as evidence of transcription (e.g. candidates located in introns)
- ▶ But it can also indicate that the candidate sequence has a different biological role incompatible with being a miRNA (e.g. exon, other ncRNA, repeat)

### Number of candidates with viable annotation

<i>Drosophila melanogaster</i>	40 582
<i>Anopheles gambiae</i>	44 210

- ▶ Most remaining candidates are annotated as intergenic regions
- ▶ Further reduction in the nr of candidates requires transcription data or that we restrict our search to miRNA genomic cluster members

### CRAVELA Framework

CRAVELA Framework	
Transcriptional potential	
Annotation	
<ul style="list-style-type: none"><li>▶ Annotation data provide information that can be used as evidence of transcription (e.g. candidates located in introns)</li><li>▶ But it can also indicate that the candidate sequence has a different biological role incompatible with being a miRNA (e.g. exon, other ncRNA, repeat)</li></ul>	
Number of candidates with viable annotation	
<i>Drosophila melanogaster</i>	40 582
<i>Anopheles gambiae</i>	44 210
<ul style="list-style-type: none"><li>▶ Most remaining candidates are annotated as intergenic regions</li><li>▶ Further reduction in the nr of candidates requires transcription data or that we restrict our search to miRNA genomic cluster members</li></ul>	

Annotation data can give a clear indication of whether a candidate is transcribed or not in the case of candidates located in introns, but most importantly, it can help exclude candidates which overlap sequences associated with other biological roles or which are annotated as repeats.

By removing candidates with non-viable annotation we reduce the number of candidates to a little over a half and the vast majority of the remaining candidates are annotated as being located in intergenic regions.

In the absence of transcriptional data and in order to further reduce the number of candidates, we can limit our search to candidates which have the potential of being part of miRNA gene clusters together with known pre-miRNAs.

- ▶ Metazoan miRNAs frequently occur in genomic clusters which may be co-transcribed
- ▶ These clusters can span up to 50 kb

<i>Drosophila melanogaster</i>	1 604
<i>Anopheles gambiae</i>	439

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡

Clusters can span up to 50 kb although generally these distances are much shorter

Even with this liberal definition of miRNA gene cluster, we dramatically reduce the number of candidates to a size that can be subjected to experimental verification or more detailed computational analysis.

## Part V

### Conclusions and Perspectives

# Conclusions and Perspectives

## Conclusions

- ▶ A comprehensive **single-genome** approach to miRNA gene finding
- ▶ The intrinsic measures introduce a **scoring scheme** which is based on general properties of pre-miRNAs alone
- ▶ The structural analysis using MinDist provides an easy-to-interpret approach to characterising the folding space of stem-loop candidates
- ▶ Many genomic stem-loops seem to satisfy all requirements for miRNA precursors
  - ▶ Represent **extensive pool** of possible pre-miRNAs
  - ▶ What determines whether they are precursors is **efficient transcription**?

## Finding pre-miRNAs

2011-06-06

## Conclusions and Perspectives

## Conclusions and Perspectives

- A comprehensive **single-genome** approach to miRNA gene finding
- The intrinsic measures introduce a **scoring scheme** which is based on general properties of pre-miRNAs alone
- The structural analysis using MinDist provides an easy-to-interpret approach to characterising the folding space of stem-loop candidates
- Many genomic stem-loops seem to satisfy all requirements for miRNA precursors
  - Represent **extensive pool** of possible pre-miRNAs
  - What determines whether they are precursors is **efficient transcription**?

We have presented a single-genome approach to miRNA gene finding.

The intrinsic measures introduce a scoring scheme is solely based on general properties of miRNA precursors

The structural analysis using the MinDist procedure provides an easy-to-interpret approach to characterising the folding space of stem-loop candidates and the identify the region of interest.

The vast quantity of stem-loops in the genome that seem to satisfy stability, robustness and structural requirements may hint at the existence of a large pool of potential miRNA precursors, and the transcription efficiency may be a critical factor deciding whether they are integrated in the miRNA processing pathway

# Conclusions and Perspectives

## Improvements

- ▶ The calculation of the *cscore* for the intrinsic measures should take into account genome heterogeneity (isochores)
- ▶ The constraints on the intrinsic measures and on the structural properties could also depend on
  - ▶ Intronic vs non-intronic miRNAs
  - ▶ miRNAs overlapping protein-coding genes in the opposite strand
  - ▶ miRNAs located in repetitive regions

## Finding pre-miRNAs

2011-06-06

## Conclusions and Perspectives

### Conclusions and Perspectives

#### Improvements

- ▶ The calculation of the *cscore* for the intrinsic measures should take into account genome heterogeneity (isochores)
- ▶ The constraints on the intrinsic measures and on the structural properties could also depend on
  - ▶ Intronic vs non-intronic miRNAs
  - ▶ miRNAs overlapping protein-coding genes in the opposite strand
  - ▶ miRNAs located in repetitive regions

Several improvements can be considered to the work we have just presented

The calculation of the combined score should take into account genome heterogeneity and calculate different background empirical distributions for each genome segment with similar characteristics.

The calculation of the intrinsic measures as well as the analysis of structural properties should also take into account whether a candidate is located in an intron or not, whether they overlap protein coding sequences in the opposite strand (and to consider their particular sequence constraints), and whether the candidate is located in repetitive regions.

## Conclusions and Perspectives

## Perspectives

- ▶ Incorporation of target prediction as a means to further reduce candidates
- ▶ Development of an evolutionary model from the analysis of the wealth of genomic stem-loops of more than one species
- ▶ Identification of miRNA regulatory modules

### Finding pre-miRNAs

2011-06-06

## Conclusions and Perspectives

Besides the readily identifiable improvements, this tool has the ability to evolve and incorporate additional evaluation strategies including the prediction of potential target genes, the development of a true evolutionary model for this structures that allows for a meaningful cross-species analysis and, eventually, the identification of miRNA regulatory modules

## Conclusions and Perspectives

## Perspectives

- Incorporation of target prediction as a means to further reduce candidates
- Development of an evolutionary model from the analysis of the wealth of genomic stem-loops of more than one species
- Identification of miRNA regulatory modules



thanks

# Thank You!