

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



Part I

Motivation

The Dark Matter of the Genome

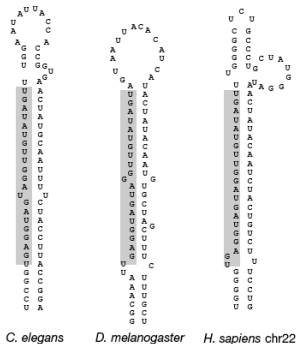
- ▶ The transcribed portion of the genome greatly exceeds the coding sequences
- ▶ Conservative estimates for the Human genome: (Claverie, *Science*, 2005)
 - ▶ Protein Coding Transcripts → **1.2% of the euchromatic genome**
 - ▶ Transcribed Sequences → **60–70% of the genome**
- ▶ Only a small portion of these non-coding transcripts (ncRNA) has been characterized
- ▶ Non-coding genomics = Non-elephant biology

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

Part III

MicroRNA Gene Finding

Publications: N D Mendes, A T Freitas, M-F Sagot.

Current tools for the identification of miRNA genes and their targets, NAR, 2009

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

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

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

MicroRNA Gene Finding

Current Approaches

- 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

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

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

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

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

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

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

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

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

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

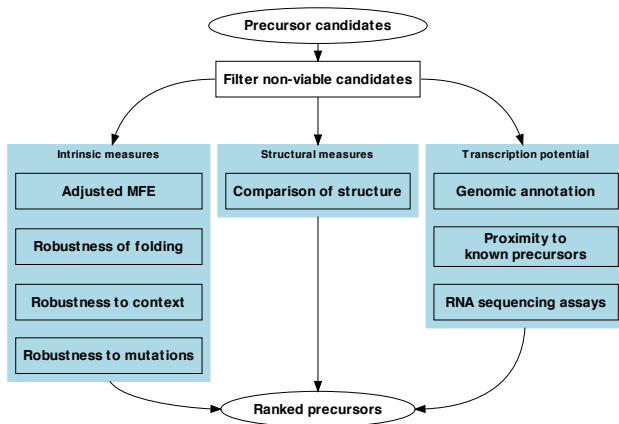
N D Mendes, S Heyne, A T Freitas, M-F Sagot, R Backofen

Navigating the unexplored seascape of pre-miRNA candidates in single-genome approaches, In preparation



CRAVELA Framework

Classification of candidates

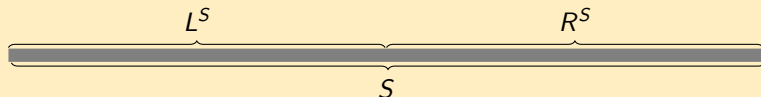


CRAVELA Framework

Extraction of candidates

Candidates obtained from a genome-wide scan

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



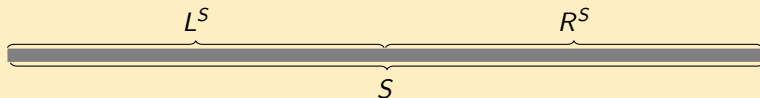
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
 - ▶ Too short ($< 16\text{bp}$) and with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

CRAVELA Framework

Extraction of candidates

Candidates obtained from a genome-wide scan

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



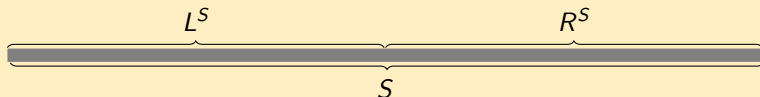
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
 - ▶ Too short ($< 16\text{bp}$) and with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

CRAVELA Framework

Extraction of candidates

Candidates obtained from a genome-wide scan

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



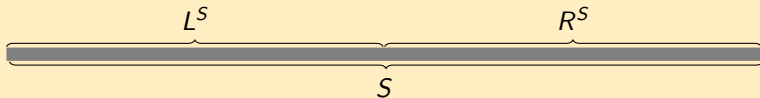
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
 - ▶ Too short ($< 16\text{bp}$) and with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

CRAVELA Framework

Extraction of candidates

Candidates obtained from a genome-wide scan

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



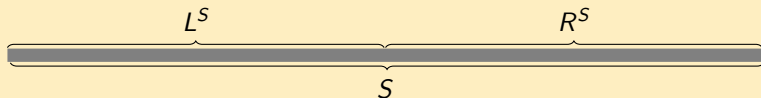
- 2 Identify candidate windows with local maxima of an evaluation function
- 3 Merge contiguous candidate windows with identical score
- 4 Fold candidate regions and excise canonical stem-loop structures found within
- 5 Eliminate unstable or aberrant precursors
 - ▶ Too short ($< 16\text{bp}$) and with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

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 ($> -20\text{kcal/mol}$)

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>)

CRAVELA Framework

Assessing the performance of intrinsic measures

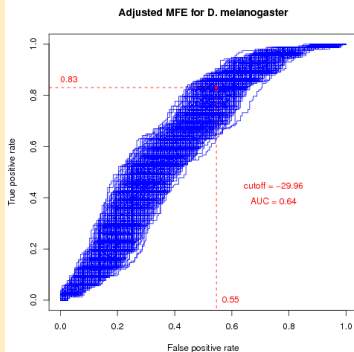
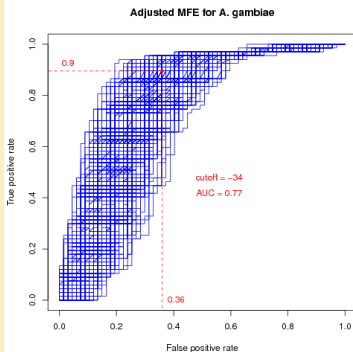
- ▶ Undersampling procedure in order to produce 1000 samples with an identical number of known precursors (constant) and non-overlapping candidates
- ▶ ROC curve for each sample reflecting the trade-off between specificity/sensitivity with respect to cutoff level
- ▶ Optimal cutoff determined using the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

CRAVELA Framework

Intrinsic Measures

Adjusted MFE (Zhang et al., 2006)

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

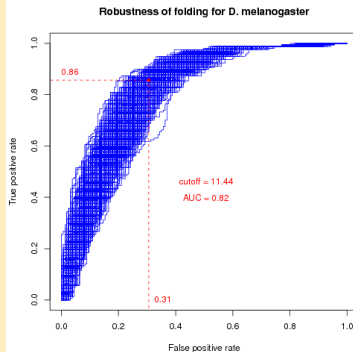
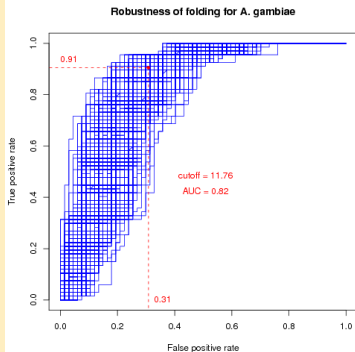


CRAVELA Framework

Intrinsic Measures

Robustness of folding

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

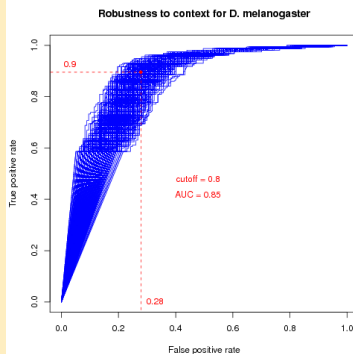
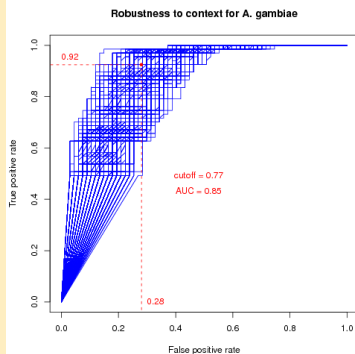


CRAVELA Framework

Intrinsic Measures

Robustness to context (Lee et al., 2008)

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

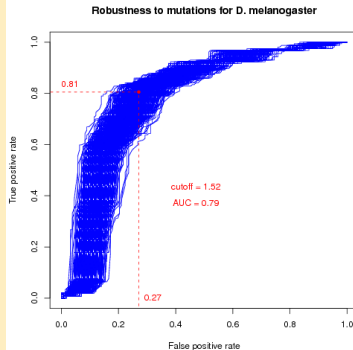
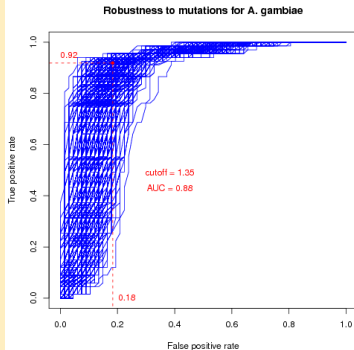


CRAVELA Framework

Intrinsic Measures

Robustness to mutations (Borenstein, 2006)

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



CRAVELA Framework

Intrinsic Measures

Composite score (*cscore*)

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

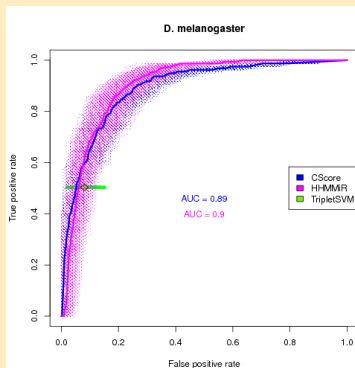
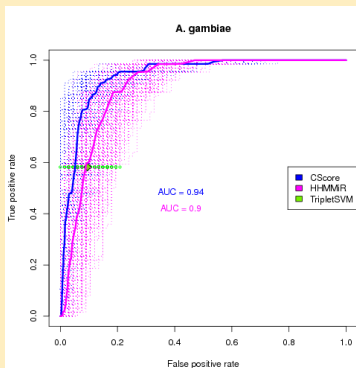
F_i is the empirical cpf for the i -th evaluation measure on a randomised genome with the same statistical properties of the genome of interest

CRAVELA Framework

Intrinsic Measures

Composite score (*cscore*)

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

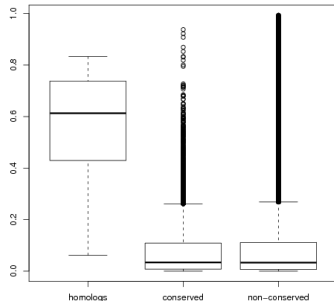


CRAVELA Framework

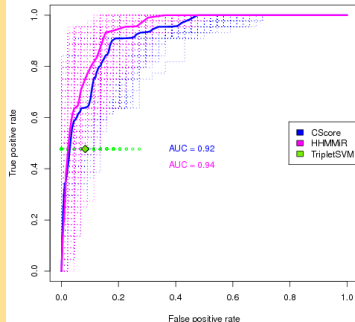
Intrinsic Measures

Exploration of candidates in *A. darlingi*

Score distribution for *A. darlingi* candidates



A. darlingi



CRAVELA Framework

Intrinsic measures

Number of candidates above cut-off

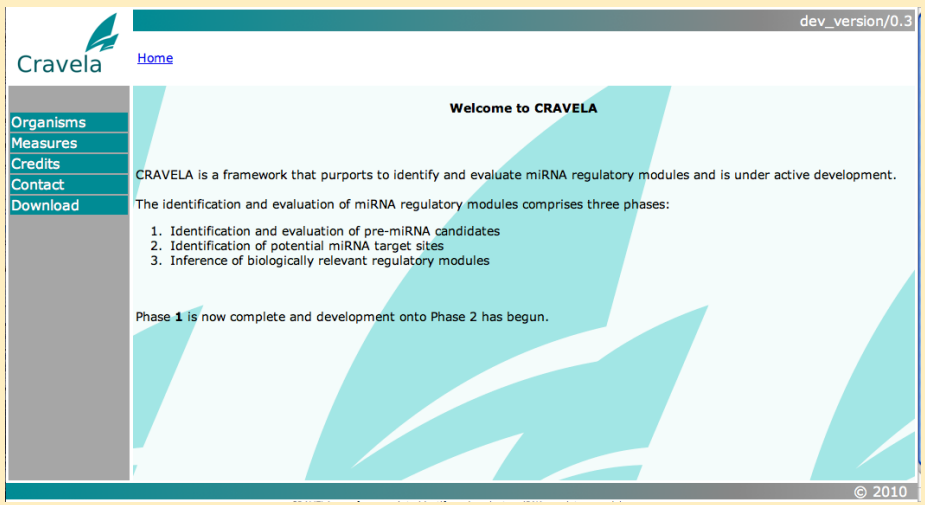
<i>Drosophila melanogaster</i>	240 751
<i>Anopheles gambiae</i>	328 829
<i>Anopheles darlingi</i>	305 681

	Cut-off	Sensitivity	Specificity
<i>Drosophila melanogaster</i>	0.30	0.83	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



The screenshot shows the CRAVELA web application interface. At the top left is the CRAVELA logo, a stylized green leaf. To its right is a dark teal header bar with the text "dev_version/0.3" in white. Below the logo is a vertical navigation menu with links: "Home" (in blue), "Organisms", "Measures", "Credits", "Contact", and "Download". The main content area has a light blue background with abstract green leaf shapes. It features a "Welcome to CRAVELA" heading, a paragraph stating the framework's purpose, a list of three phases, and a status update for Phase 1. A copyright notice "© 2010" is in the bottom right corner.

dev_version/0.3

CRAVELA

[Home](#)

Organisms

Measures

Credits

Contact

Download

Welcome to CRAVELA

CRAVELA is a framework that purports to identify and evaluate miRNA regulatory modules and is under active development.

The identification and evaluation of miRNA regulatory modules comprises three phases:

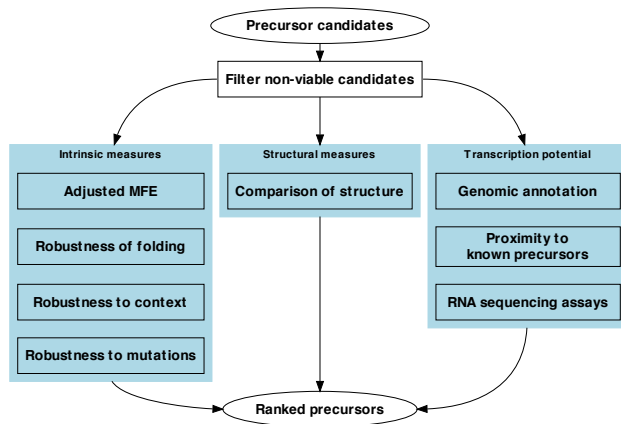
1. Identification and evaluation of pre-miRNA candidates
2. Identification of potential miRNA target sites
3. Inference of biologically relevant regulatory modules

Phase **1** is now complete and development onto Phase 2 has begun.

© 2010

CRAVELA Framework

Classification of candidates



CRAVELA Framework

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

Structural analysis

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

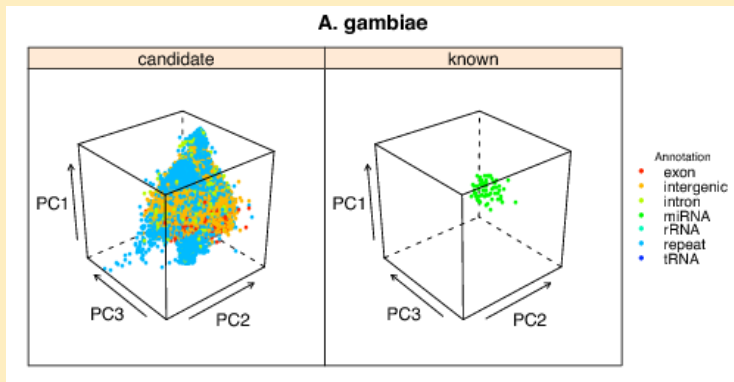
Feature space relative positions reflect structural similarity

- ▶ 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

CRAVELA Framework

Structural analysis

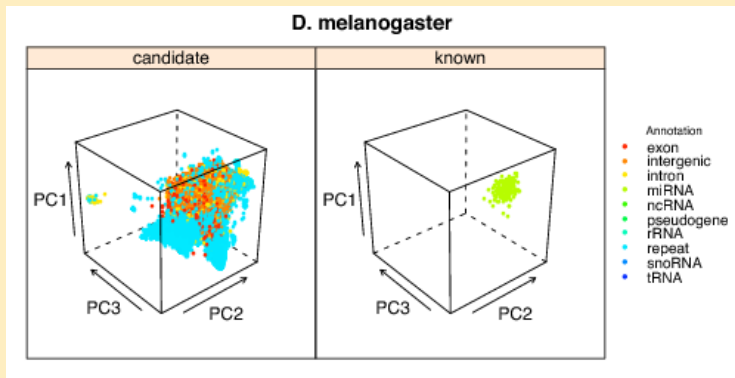
Feature space (3D)



CRAVELA Framework

Structural analysis

Feature space (3D)



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.

CRAVELA Framework

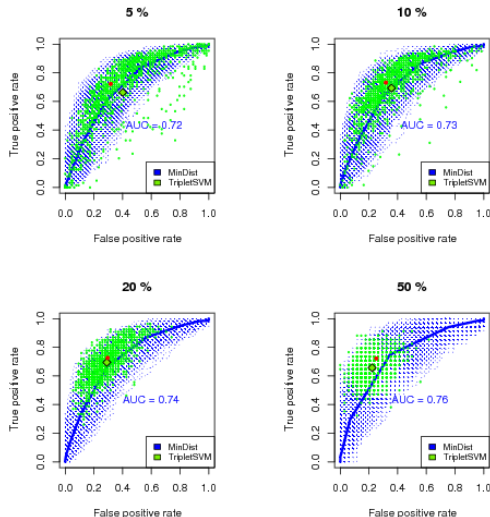
Assessing the performance of MinDist

- ▶ 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

CRAVELA Framework

Classification of candidates

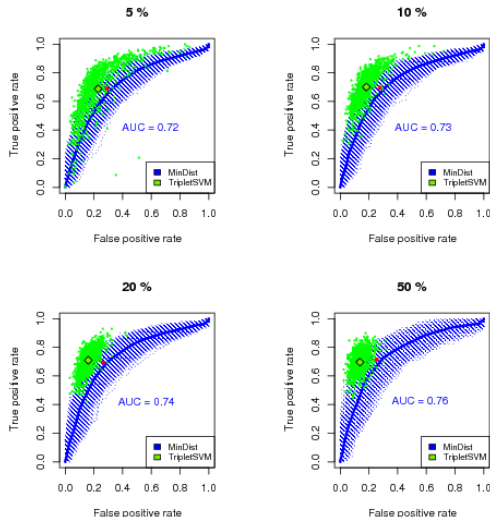
A. gambiae



CRAVELA Framework

Classification of candidates

D. melanogaster



CRAVELA Framework

Classification of candidates

% 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 $\left(2 \frac{TP/(TP+FP) * TPR}{TP/(TP+FP) + TPR}\right)$ 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*

CRAVELA Framework

Structural analysis

Estimating optimal cut-off

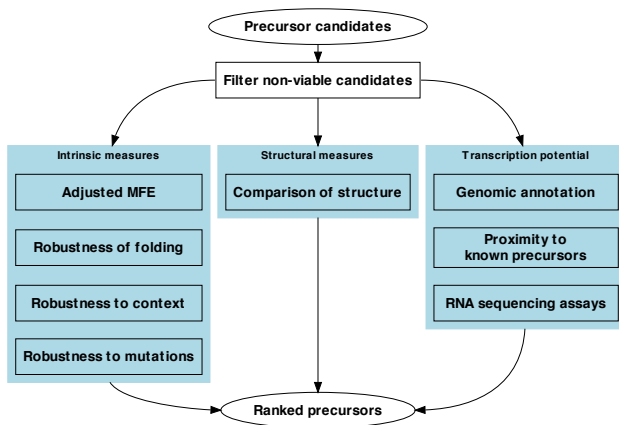
- ▶ Each sample group average optimal cut-off and proportion of known precursors used follow a log-linear law until reaching about 90%
- ▶ Distance used to calculate acceptance region estimated by extrapolating the log-linear law to 100%

Number of candidates within acceptance region

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

CRAVELA Framework

Classification of candidates



CRAVELA Framework

Transcriptional potential

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

Transcriptional potential

miRNA genomic clusters

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

Number of candidates up to 50 kb away from known precursors

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

Experimental testing or further detailed structural analysis can be carried out

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**?

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

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

Thank You!