Hairpins in a Haystack

Jana Hertel

Bioinformatics Group,
Department of Computer Science and Interdisciplinary Center for Bioinformatics,
University Leipzig

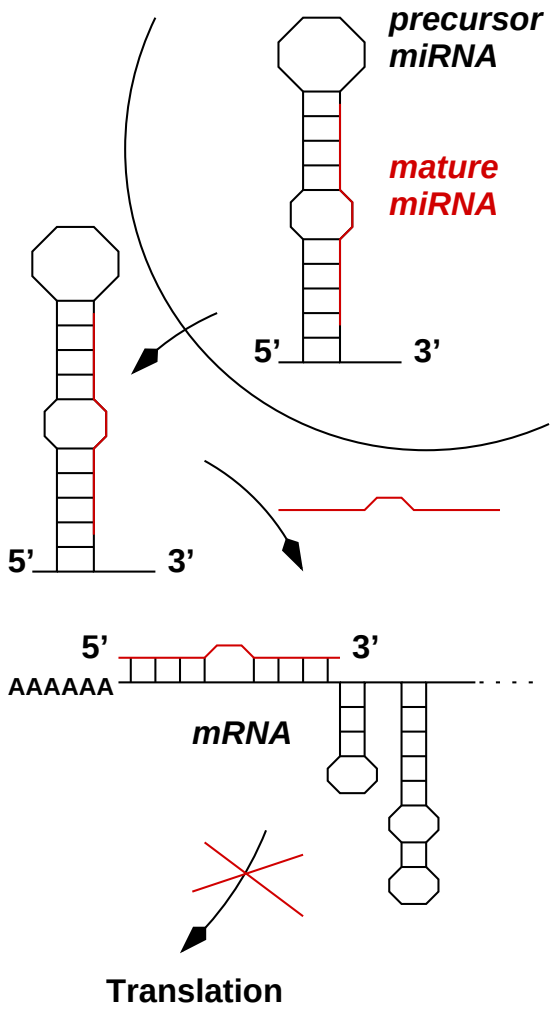
Leipzig, February 22, 2006

Outline

- **Introduction**
  - Background
  - Detection
  - Basis
  - Purpose
- **Methods**
  - Hairpin filter
  - Descriptors and SVM
  - SVM training
- **Application**
  - Homo sapiens
  - Nematodes
  - Seasquirts
- **Discussion**

MicroRNAs - Background

- class of noncoding RNAs
- important regulatory functions
- longer transcripts (pre-miRNAs) $\rightarrow \sim$ 100nt
- functional mature miRNA in one stem side $\rightarrow \sim$ 22nt
- mature miRNA highly conserved
- bind to 3'UTRs of mRNA targets
  - supress expression
  - mark for degradation



MicroRNAs - Detection

- candidates homologous to miRNAs
- candidates adjacent to known miRNAs

Problem:

- candidates that do not feature these facts can not be found

MicroRNAs - Detection

- several approaches for detecting novel miRNA genes
- secondary structure, 3' and 5' patterns in stem loop

Examples:

- `miRscan`[1] (nematodes), `miRseeker`[2] (insects), `miralign`[3] (vertebrates)
- candidate search, classifying by features partly machine learning

[1] Lim *et al.* 2003
[2] Lai *et al.* 2003
[3] Wang *et al.* 2005

Basis of this approach

- genome-wide screens for ncRNAs
- RNAz[4] $\rightarrow$ evolutionary conserved secondary structure in multiple sequence alignments
- automatic tools neccessary to assign candidates to ncRNA classes

[4] Washietl *et al.* 2005

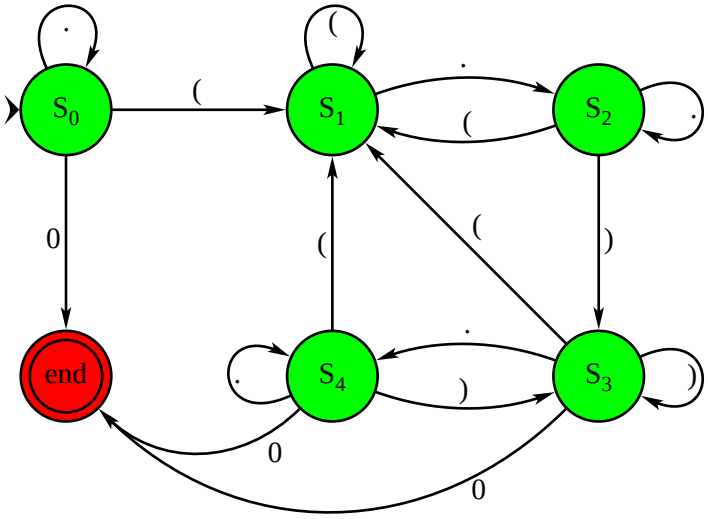
Purpose of RNAmicro

- RNAmicro works as ”sub-screen” for ncRNA surveys
- SVM based tool for classifying miRNAs among ncRNAs

Paritcular steps

1. detect *almost* hairpins
2. computation of descriptors
3. SVM classification

Detecting *almost* hairpins



- specify window of alignment
- consensus sequence and structure
- dot-parantheses string read by automaton
- start and length of each stem loop stored
- accept structure if
  - exactly 1 stem loop $>$ 10nt
  - other smaller stem loop $<=$ 4 nt

Computation of Descriptors

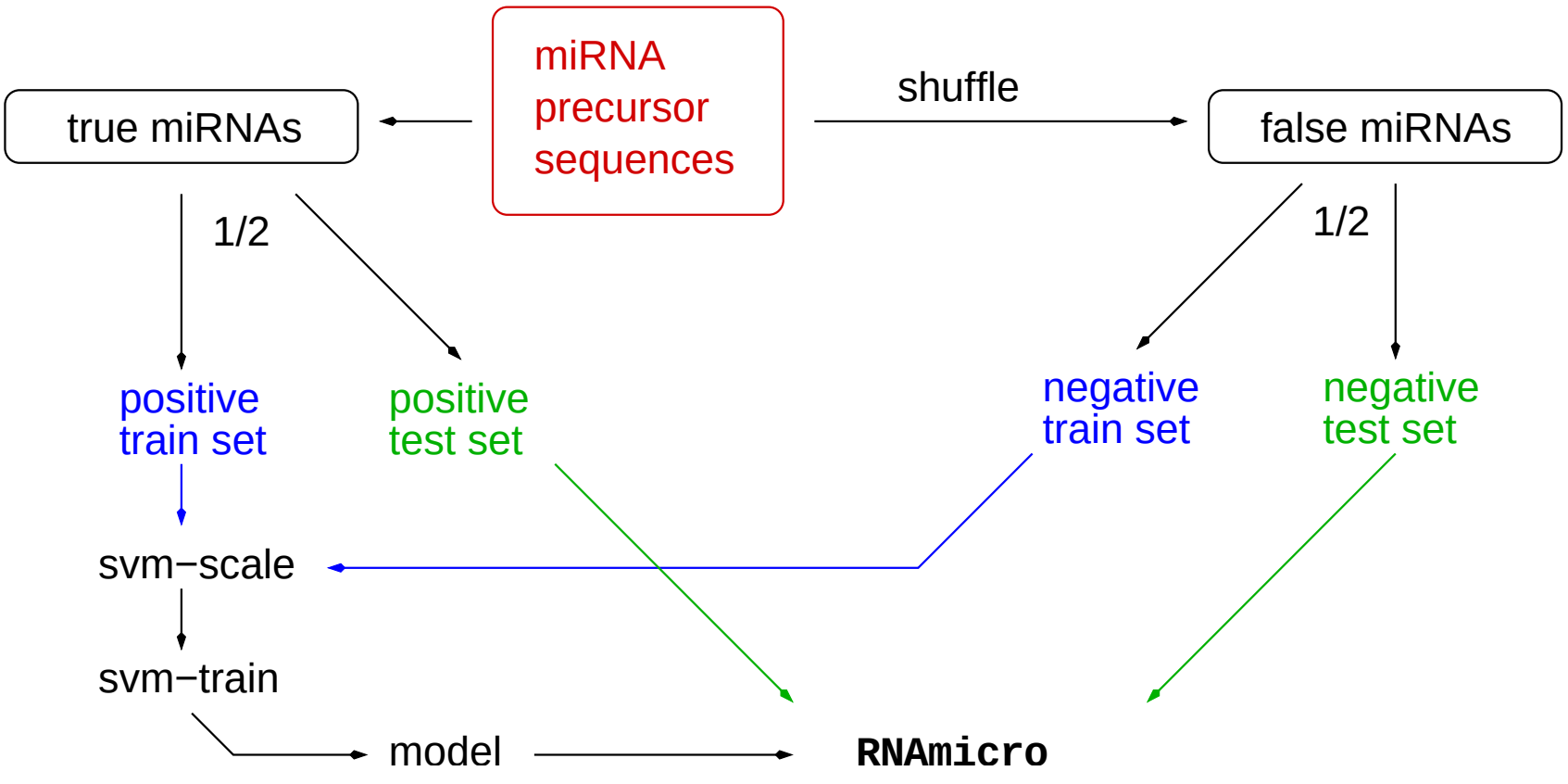
stem length	z-score	consensus mfe
loop length	mean single mfe	3’/5’ stem, loop entropy
G+C content	adjusted mfe	23nt block entropy
	mfe index	

SVM implementation

- SVM from `libsvm`[5]
- scale descriptors to $[-1, +1]$
- model: rbf kernel, probability estimates

[5] Chang and Lin, 2001

Initial training



Results of initial training

- 134/147 (90%) sensitivity, 381/383 (99%) specificity
- train SVM again with entire datasets
- test on RNAz screens of nematodes and seasquirts
  $\rightarrow$ significant number of known ncRNAs false classified
  $\rightarrow$ initial negative set not sufficiently good

Retraining the SVM

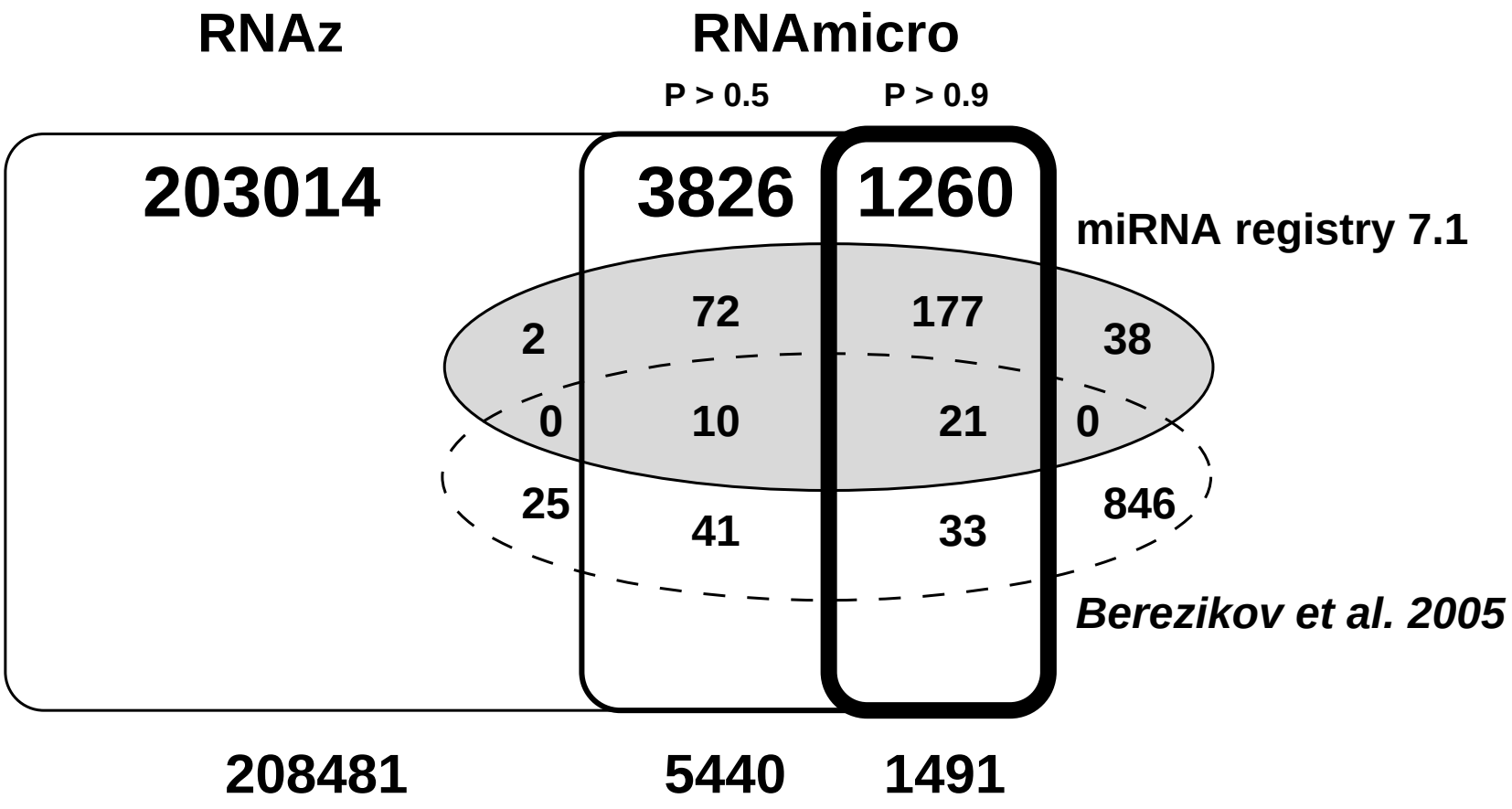
- ncRNA alignments extracted from `Rfam` database
- add known false positives to negative train set
- iterate process of adding false positives and retraining until no significant improvement on the `Rfam` dataset

Data

- vertebrate genomes[6], nematode and urochordate[7] ncRNA alignments
- screen with 70,100 and 130 nt window
- retaining best (p!) non-overlapping hits of each alignment

[6] Washietl *et al.* 2005
[7] Missal *et al.* 2005

General results

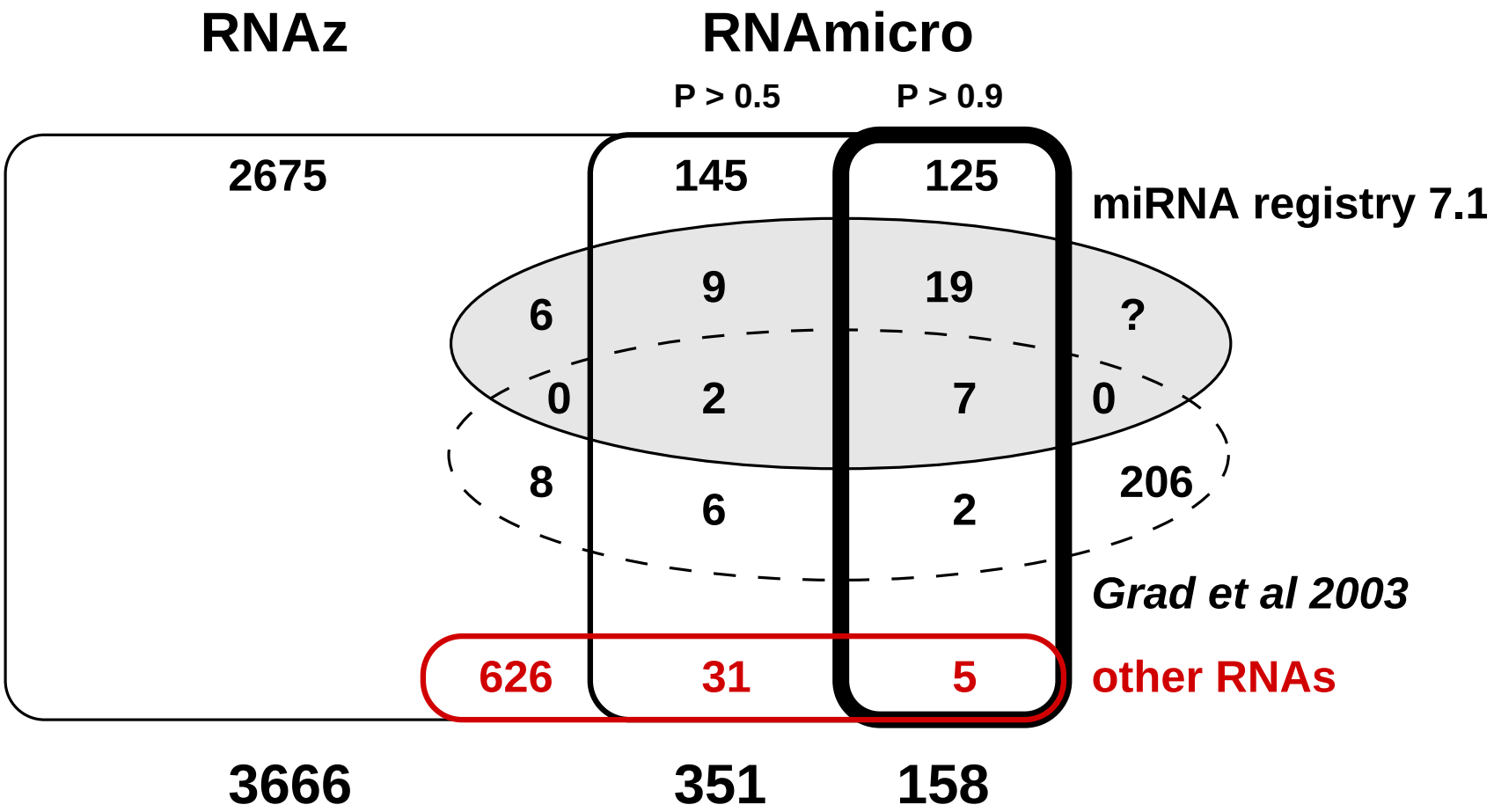


Verification

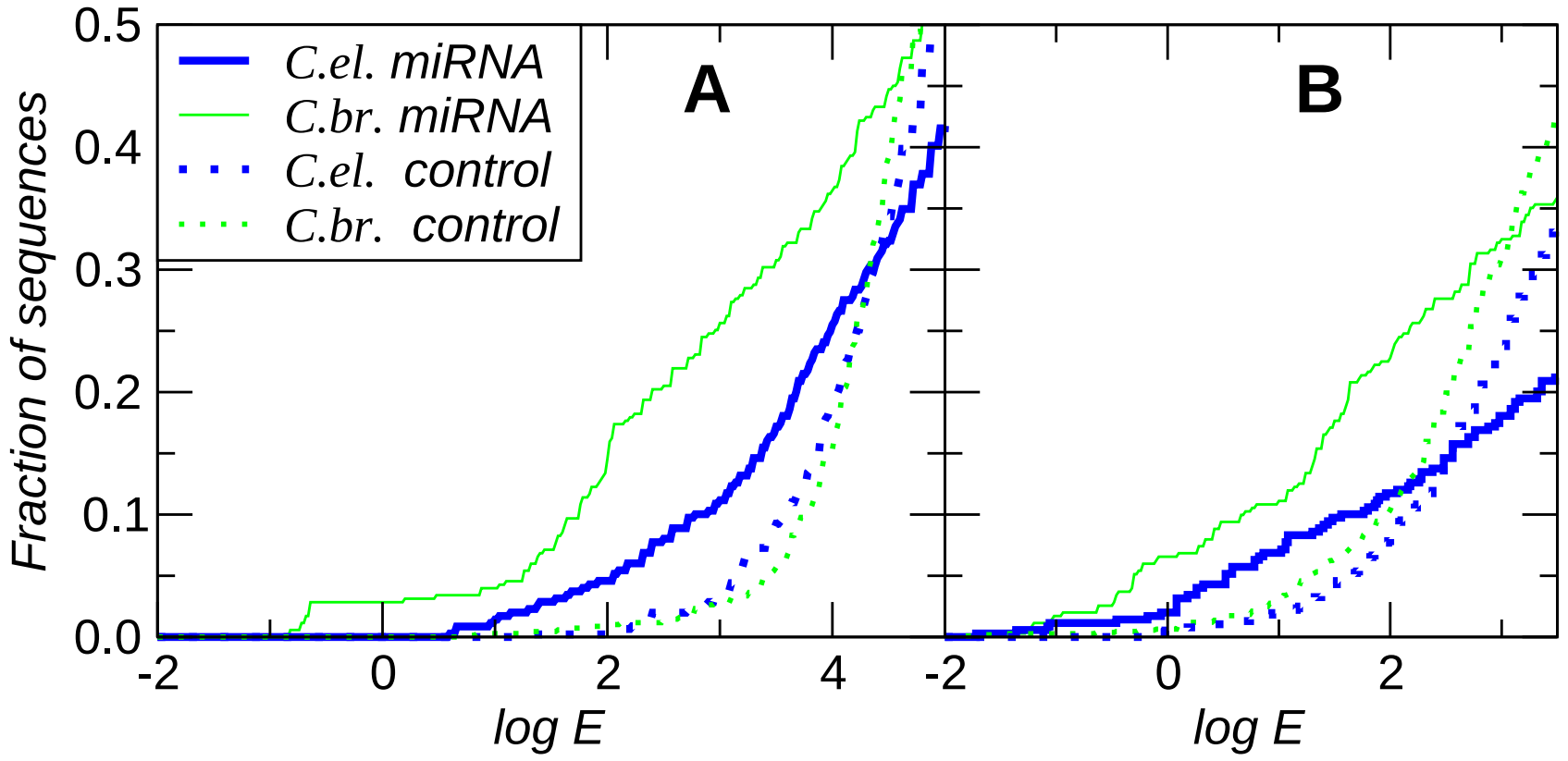
- re-evaluate `RNAmicro` candidates with other SVM approach[8]
- very restrictive hairpin filter
  - 3077 / 5440 with P $>$ 0.5 passed fi lter, 1590 recognized
  - 953 / 1481 with P $>$ 0.9 passed fi lter, 657 recognized
- 4245 / 5440 candidates not associated with protein coding genes
- 1107 candidates located within introns (36 known)

[8] Xue *et al.* 2005

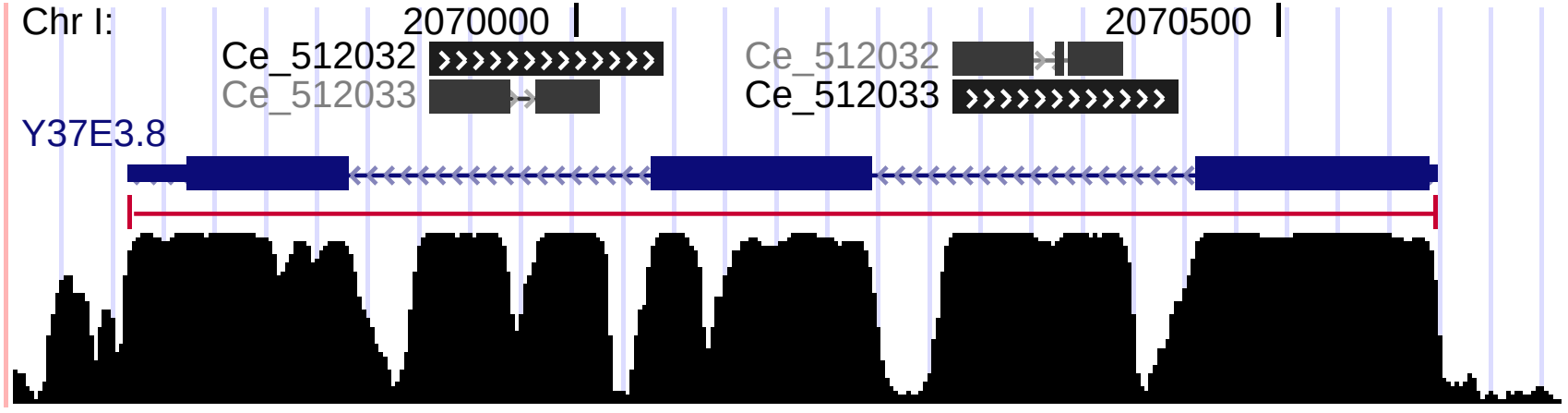
General results



Upstream motif

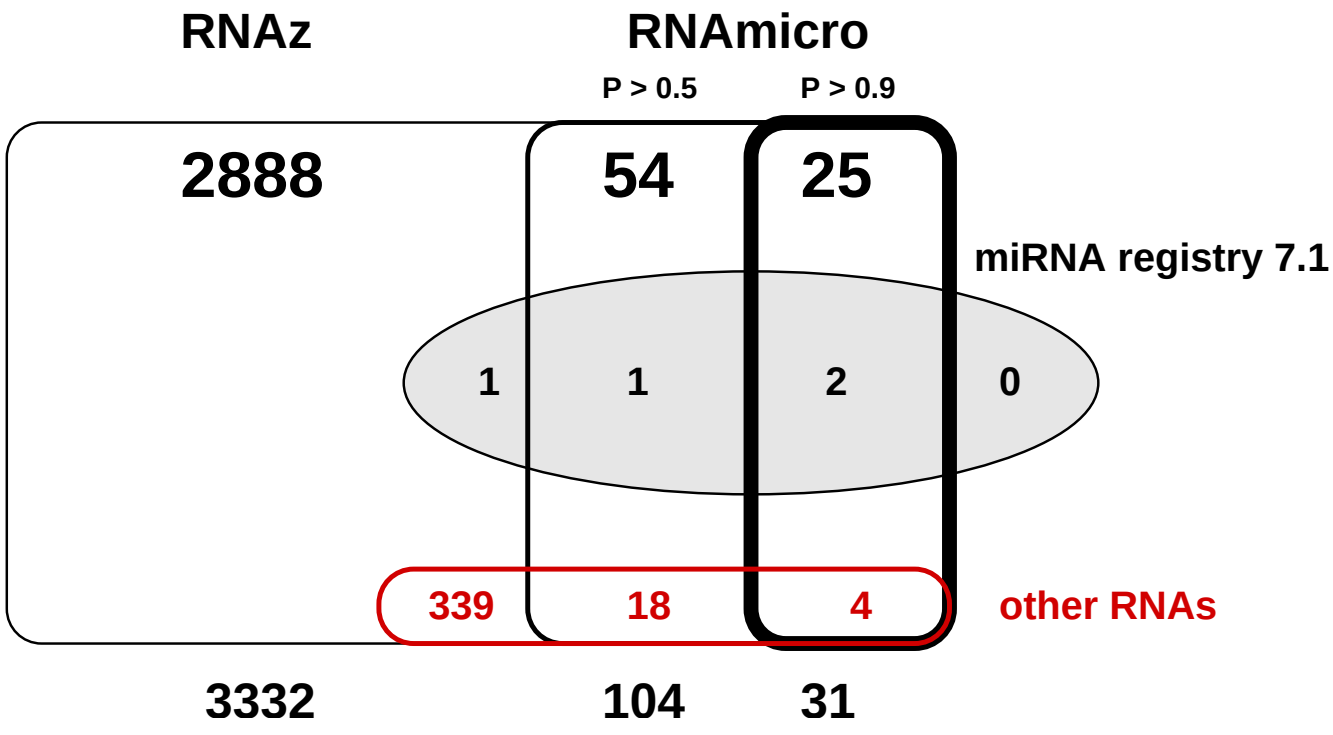


Intronic and clustered miRNAs



- 176 / 351 intronic candidates
- 30 clusters with 131 members

General results



- 5 clusters with 10 members

Summary

- `RNAmicro` designed for classification of ncRNA alignments
- applied to 3 recent `RNAz` based studies
- large number of novel miRNA candidates
- verification through
  - comparison with other approaches and annotations
  - anaylising genomic location to other candidates
  - location in introns
- large number of `RNAmicro` predictions correspond to real miRNAs
- only small fraction of true miRNA repertoire known