# Identification of snoRNAs using Machine Learning - Improving the snoReport

João Victor de Araujo Oliveira[1], Pedro A. Berger[1], Peter Stadler[2], Maria Emília M. T. Walter[1] and Jana Hertel[2]

[1]Department of Computer Science, University of Brasilia, Brazil

[2]Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

February 21, 2015

# Outline

# Small Nucleolar RNA (snoRNA)

- snoRNAs are 60 to 300nt ncRNA that accumulate in the nucleolus
- Function: Chemical modifications in rRNAs, tRNAs e snRNAs
- Classified based on their characteristic sequence elements (called boxes) in two classes:
  - C/D box *snoRNA*
  - H/ACA box *snoRNA*

# C/D box snoRNA

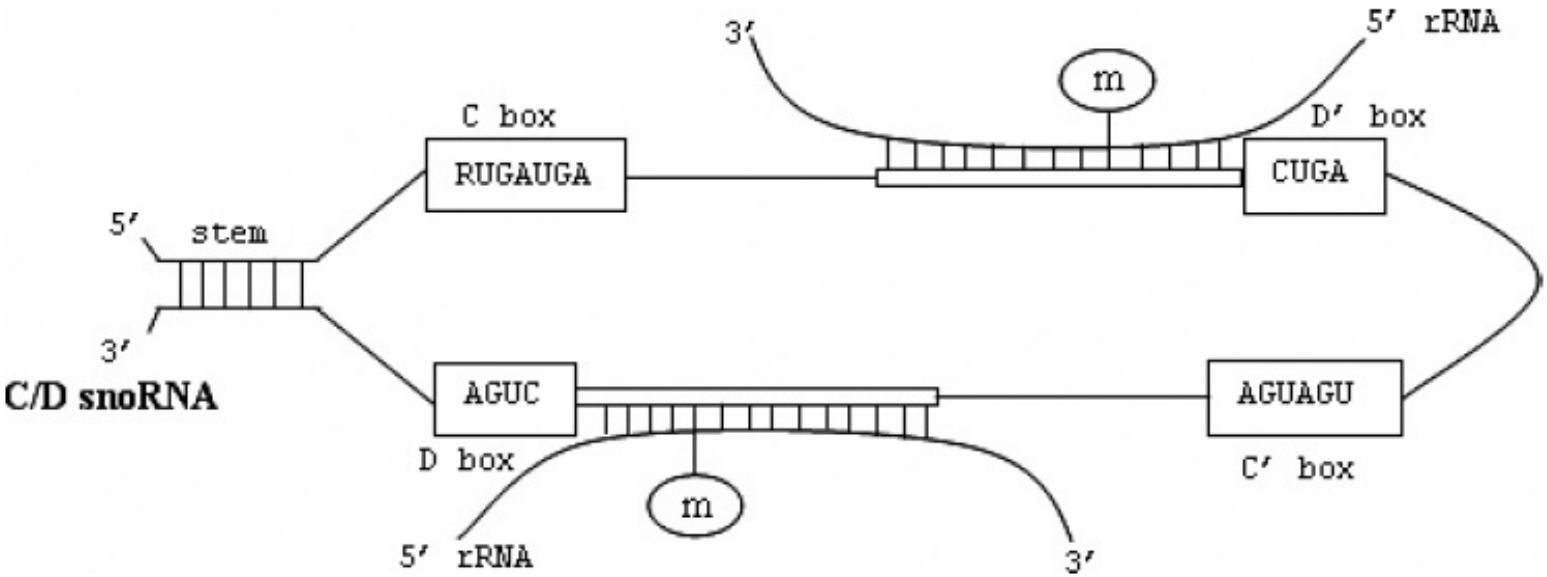



Figure: Canonic representation of C/D box snoRNA (Yang et al., 2006)

# H/ACA box *snoRNA*

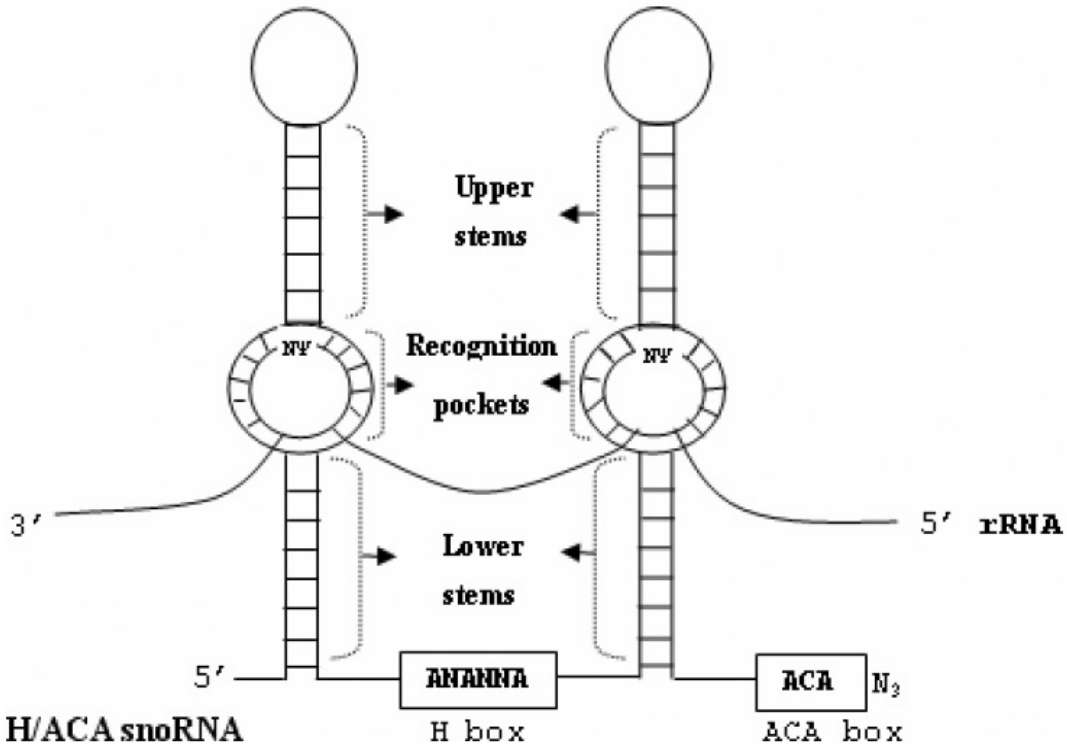



Figure: Canonic representation of H/ACA box snoRNA (Yang et al., 2006)

# Machine Learning

- Sub-area of Artificial Intelligence
- *"A computer program is said to* **Learn** *from Experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"* (Mitchell, 1997)
- Examples of Algorithms:
  - Support Vector Machine (SVM)
  - Ensemble Methods like Random forest
  - Many others...

# snoReport (Hertel et al., 2008)

- Tool that identifies the two main classes of snoRNAs:
  - C/D box *snoRNA*
  - H/ACA box *snoRNA*
- Method:
  - Combination of secondary structure prediction and machine learning (SVM)
  - It had been trained on almost exclusively mammalian sequences
  - It incorporates default meta-parameters for the SVM classifier

# snoReport

- It does not use information of putative target sites within ribosomal or spliceosomal RNA
    - This information can dramatically improve identification sensitivity and specificity
- Orphan snoRNAs are being discovered:
    - With no known target or
    - Target specific mRNA
    - Suggest other functions
- snoReport can detect both orphan snoRNAs and guide snoRNAs

# Problem

snoReport needs refinements because nowadays we have:

- Many new sequences of snoRNAs for different vertebrate organisms
- New versions of all tools used in snoReport previously (Vienna RNA Package and libSVM)

# New snoReport

- This work intends to build a new snoReport:
  - Extracting different features for both C/D and H/ACA box snoRNAs
  - A more sophisticated technique in the SVM training phase:
    - Recent data from vertebrate organisms
    - Different approach to refine SVM meta-parameters
    - Using different machine learning algorithms in order to compare performance
  - Using new versions of the tools and databases previously used to build snoReport

# New snoReport

- Other goals:
  - Build models specific to different classes of organisms (e.g. plants, archaea, animals, fungi...)
  - Parallelize certain functions in snoReport, in order to improve performance
  - Give the opportunity to the user builds his own model from his data
  - Apply our program to a set of previously predicted sequences in humans, nematodes, drosophilids, platypus,chickens, leishmania and others
  - Try to find new snoRNAs in specific organisms, like the fungus *Paracoccidioides brasiliensis*

# PWMs Generator

- Position-specific weight Matrices (PWMs) were used to represent each characteristic sequence motif of both classes of snoRNA
- PWMs obtained by scanning the boxes of known human and yeast snoRNAs
- For each nucleotide, the probability found in a position of the box motif was calculated

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & U \end{array} \\ \begin{pmatrix} 0.1250 & 0.6250 & 0.1250 & 0.1250 \\ 0.1250 & 0.1250 & 0.1250 & 0.6250 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.8750 & 0.1250 & 0.0000 & 0.0000 \end{pmatrix} \end{array}$$

Figure: PWM of box D (CUGA)

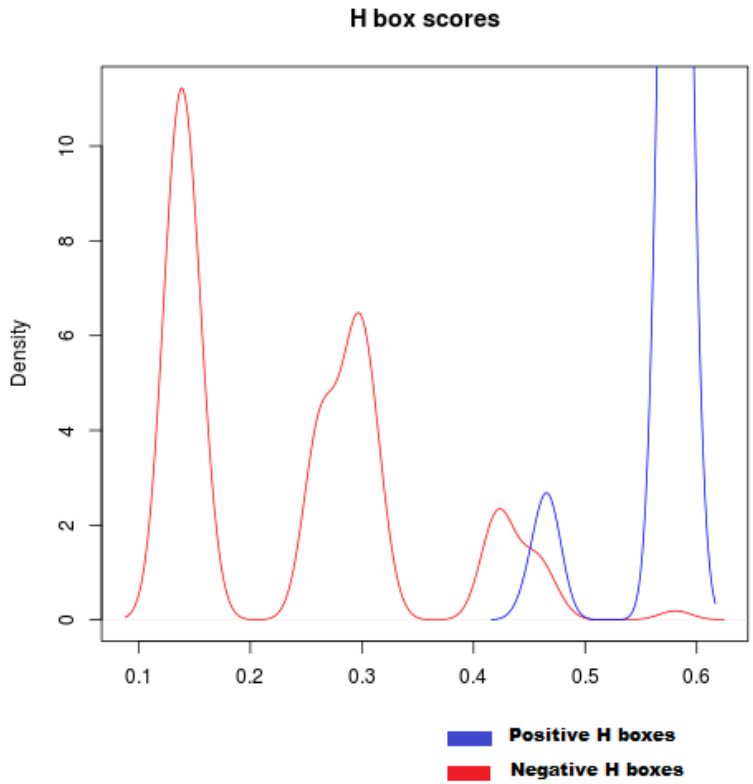
# Threshold Analysis

- We scanned known snoRNAs sequences and calculated a PWM's-based score for each candidate of some box
- 101 known H/ACA box snoRNAs
- C/D box not scanned yet : (

>HACA_101−1_G.gallus_(adopted_5)_46649447,46649585_+_AAAGCA_62_ACA_134
AAGTCAGCTAAGTGATACTGCAGCATTATGAATCATCTCTGTAACACTGAGCTGCTTTT
TAAAAGCACAATTTTGTGGTGATAAAAGGTTTGAAGGTGACAAGCCTGCAGATCTTATT
CTACTTACCCACAGAACATAG

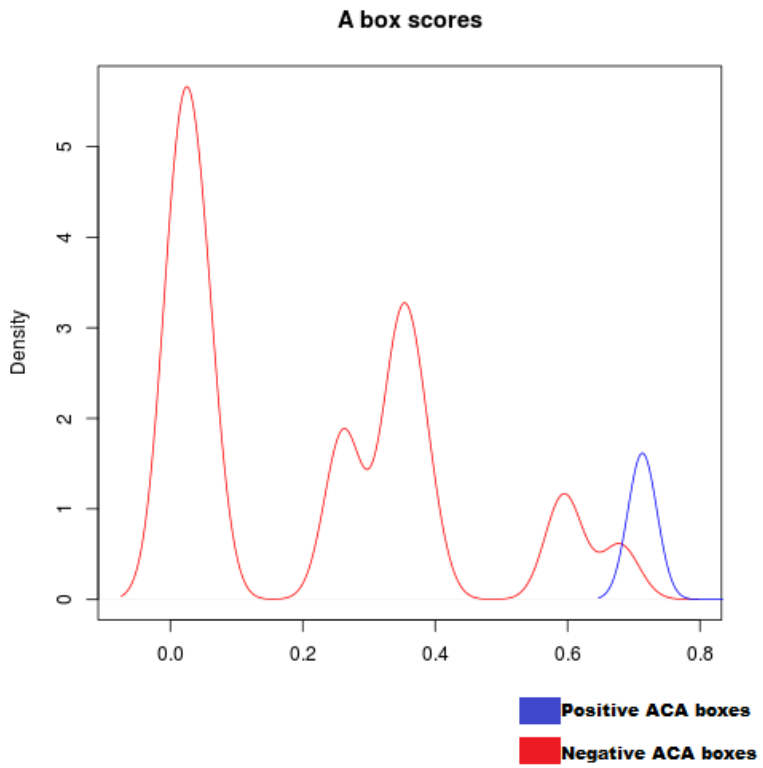
# H box threshold

- Hbox threshold = 0.4519
- Common Hbox = ANANNA
- All N's have probability of 0.25 to occur
- This threshold accepts Hbox = ANANNN, which is possible in some cases

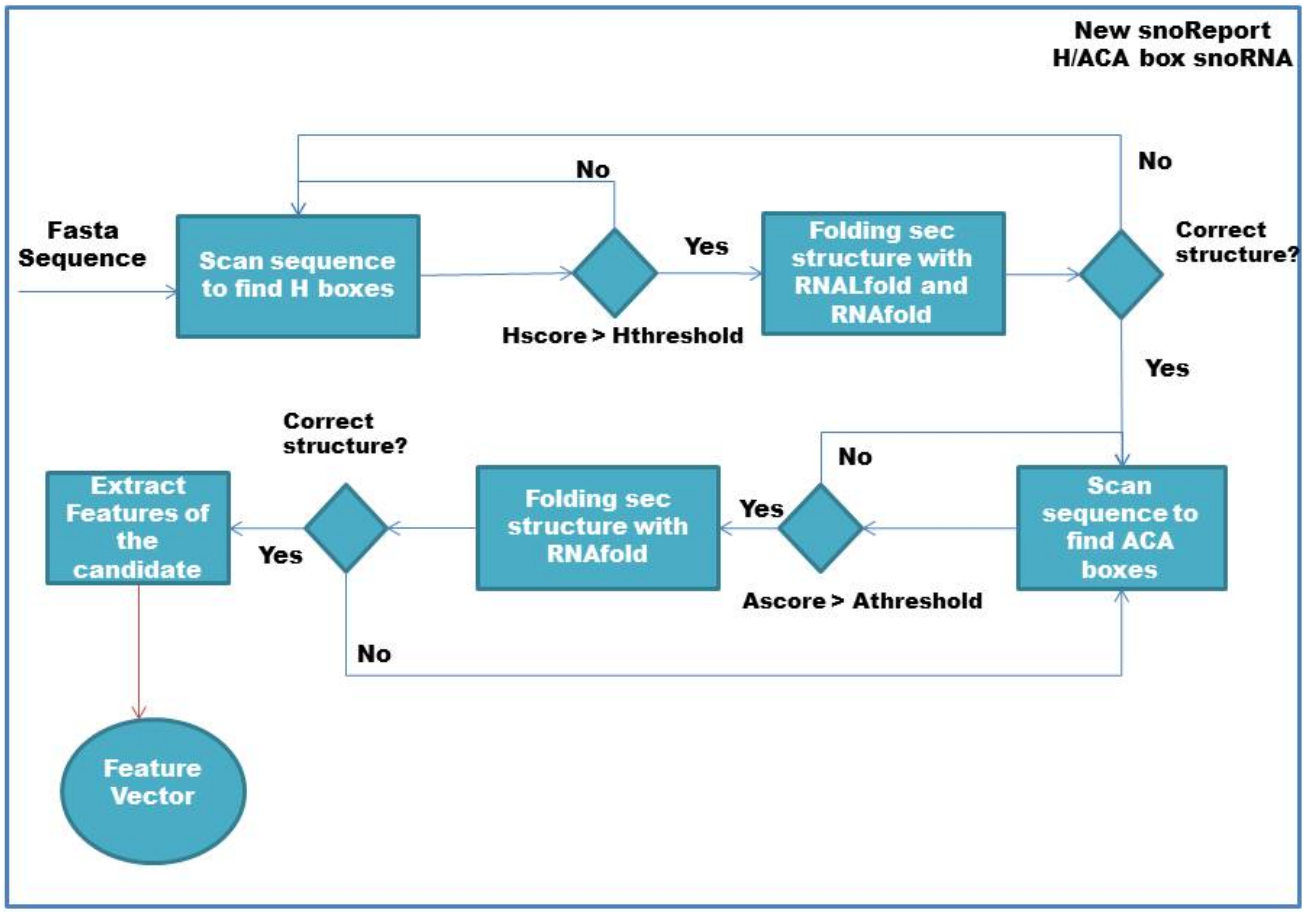


# ACA box threshold

- ACA threshold = 0.6770
- Common ACA box = ACA (obvious)
- Some cases we find ATA as ACA box
- AUA, AAA and AGA have good scores (negative samples)



# Search of the H/ACA box candidates



# Extraction of feature vectors for H/ACA box snoRNA

Table: New attributes (in blue), compared to the previous version: *AC* and *GU*, *zscore*, *Hscore* and *ACAscore*, and *LloopSC*, *RloopSC*, *LloopYC* and *RloopYC*

| | |
|---|---|
| *mfeC* | MFE of the secondary structure with restrictions in RNAfold |
| *AC*, *GU*, *GC* | AC, GU and GC content |
| *zscore* | zscore obtained by RNAz |
| *Hscore* | Score of the H box |
| *ACAscore* | Score of the ACA box |
| *LseqSize* | Number of nucleotides before the H box |
| *RseqSize* | Number of nucleotides between H and ACA box |
| *LloopSC* | Lenght of the loop, where we find the pocket region containing the target region, near to the H box |
| *RloopSC* | Length of the loop, where we find the pocket region containing the target region, more close to the ACA box |
| *LloopYC* | Symmetry of the loop containing the pocket region near to the H box |
| *RloopYC* | Symmetry of the loop containing the pocket region near to the ACA box |
| *LoopSym* | Symmetry of the loops inside the hairpin |

# Extraction of feature vectors for C/D box snoRNA

Table: New attributes (in blue), compared with the previous version, were included: $Dcd'$, $Dd'c'$ and $Dc'd$, $D'_{score}$ and $C'_{score}$

| | |
|---|---|
| $mfe$ | MFE of the secondary structure without restrictions in RNAfold |
| $mfeC$ | MFE of the secondary structure with restrictions in RNAfold |
| $E_{avg}$ | Mean of the MFE |
| $E_{stdv}$ | Standard deviation of the MFE |
| $ls$ | Length of the terminal stem |
| $Dcd$ | Distance between C and D boxes |
| $Dcd'$ | Distance between C and D' boxes |
| $Dd'c'$ | Distance between D' and C' boxes |
| $Dc'd$ | Distance between C' and D boxes |
| $C_{score}$ | score of the C box |
| $D'_{score}$ | score of the D' box |
| $C'_{score}$ | score of the C' box |
| $D_{score}$ | score of the D box |
| $GC$ | GC content |

# Training dataset

- We will use different dataset used in the previously version
  - mRNAs
  - snRNAs
  - rRNA
  - Sequences obtained applying a dinucleotide shuffling procedure in snoRNA samples

# Training and test phases

- First, we will use *subset.py* to divide randomly the samples for training and testing datasets
  - Provide a stratified selection to guarantee the same class distribution in both subsets
  - 60% for training and 40% for testing
- Second, we will use *svm − scale* to scale the feature vector from -1 to 1

# Workflow of training and test phases

- After, we will perform a grid search for the parameters $C$ and $\gamma$ using *grid.py*
  - It uses a cross validation technique (10-fold) to estimate some performance criteria of each combination of $C$ and $\gamma$
  - We investigated all combinations of these parameters ranging both from $2^{-15}$ to $2^{15}$, shifting $2^1$ for each step of the grid-search
  - This approach can be used with many different performance criteria, e.g., accuracy, F-score, BAC (balanced accuracy).

# Grid search

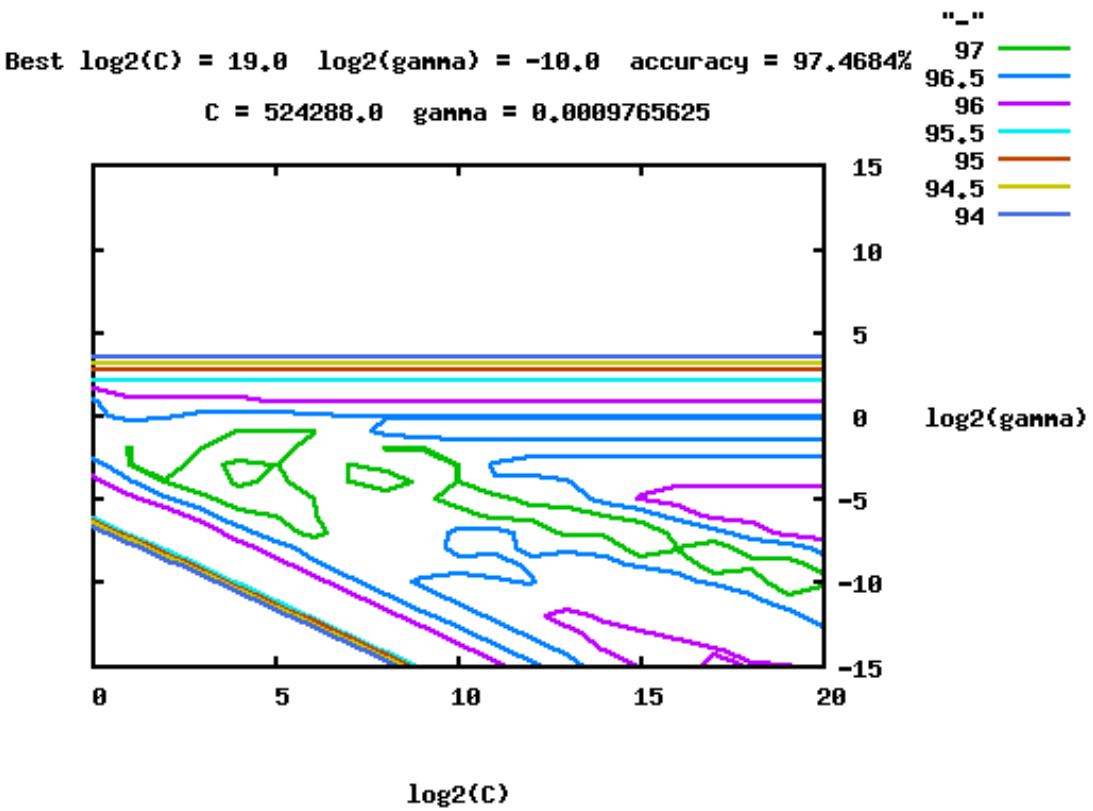
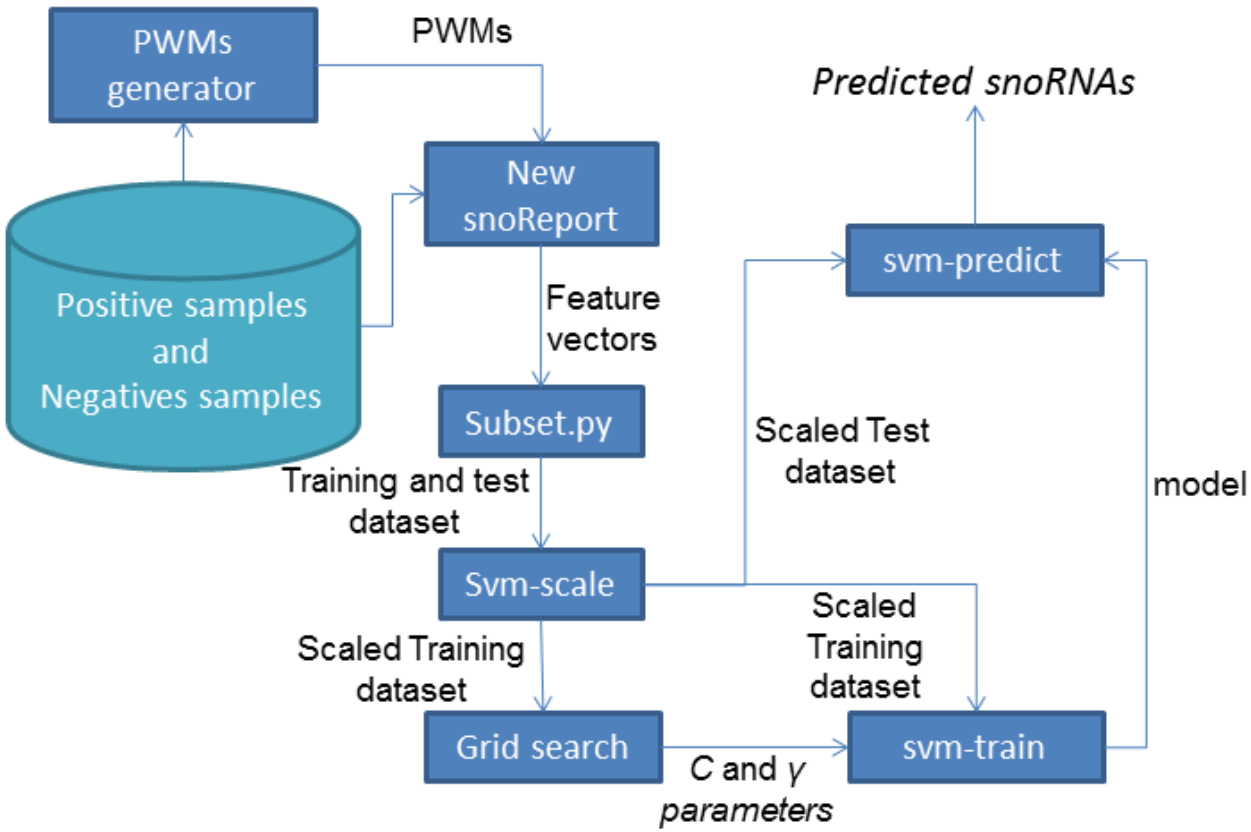



Figure: Accuracy as a criterion for H/ACA box snoRNA classification

# Training and test phases

- Training will be performed using *svm-train*, which used a SVM called C-SVM with RBF kernel and 10-fold cross validation
- *svm-train* returns a classifier (called model) used as input in *svm-predict* to predict snoRNAs from sequences not used in the training phase

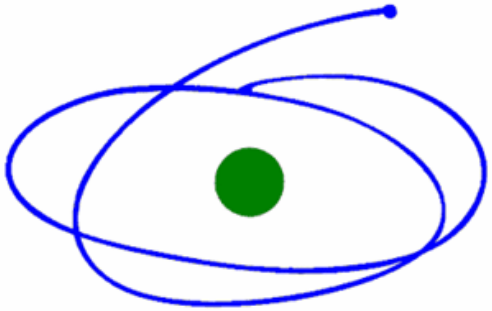
# Workflow of the new snoReport



# Acknowledgments


Universidade de Brasília

UNIVERSITÄT LEIPZIG


CAPES

Not thank you for this!


# But thank you for your attention! :)

joaovicers@gmail.com

# Performance criteria

Table: *TP* means true positive, *FN* means false negative and *prec* means precision

| | |
|---|---|
| Accuracy | $(TP + TN)/(P + N)$ |
| Precision | $TP/(TP + FP)$ |
| Sensitivity (Recall) | $TP/(TP + FN)$ |
| Specificity | $TN/(TN + FP)$ |
| F-score | $(2 * Prec * Recall)/(Prec + Recall)$ |
| BAC | $(Sensitivity + Specificity)/2$ |
| AUC | Area under the ROC curve |

# Statistics

Table: Results of the training and test phases of H/ACA box snoRNAs, using the $C$ and $\gamma$ parameters of both versions of snoReport. The improved measures are in blue

| | Criterion | snoReport | new snoReport |
|---|---|---|---|
| **Training phase** | | | |
| | Accuracy | 96.55% | 97.35% |
| | Precision | 94.29% | 82.76% |
| | Sensitivity | 54.10% | 78.69% |
| | Specificity | 99.75% | 98.76% |
| | F-score | 68.75% | 80.67% |
| | BAC | 76.92% | 88.72% |
| | AUC | 96.46% | 98.62% |
| **Test phase** | | | |
| | Accuracy | 96.72% | 97.93% |
| | Precision | 92% | 86.84% |
| | Sensitivity | 57.50% | 82.5% |
| | Specificity | 99.63% | 99.07% |
| | F-score | 70.77% | 84.62% |
| | BAC | 78.57% | 90.79% |
| | AUC | 98.67% | 98.41% |

# Statistics

Table: Results of the training and test phases of C/D box snoRNA, using the $C$ and $\gamma$ parameters of both versions of snoReport. The improved measures are in blue

| **Training phase** | Criterion | snoReport | new snoReport |
|---|---|---|---|
| | Accuracy | 95.67% | 97.84% |
| | Precision | 96.43% | 92.96% |
| | Sensitivity | 77.14% | 94.29% |
| | Specificity | 99.42% | 98.56% |
| | F-score | 85.71% | 93.62% |
| | BAC | 88.28% | 96.42% |
| | AUC | 95.67% | 99.13% |
| **Test phase** | | | |
| | Accuracy | 96.75% | 98.56% |
| | Precision | 100% | 97.67% |
| | Sensitivity | 80% | 93.33% |
| | Specificity | 100% | 99.57% |
| | F-score | 88.89% | 95.46% |
| | BAC | 90% | 96.45% |
| | AUC | 99.52% | 98.96% |

# Validation on real data

Table: Applying new snoReport into snoRNA sequences of different organisms: number of predicted candidates / number of candidates identified in the cited reference

| | | |
|---|---|---|
| **Human** (Yang et al., 2006) | C/D: 20/21 | H/ACA: 26/32 |
| **Nematodes** (Zemann et al., 2006) | C/D: 57/108 | H/ACA: 30/60 |
| **Drosophilids** (Huang et al., 2005) | C/D: 55/63 | H/ACA: 28/56 |
| **Platypus** (Schmitz et al., 2008) | C/D: 130/144 | H/ACA: 47/73 |
| **Chicken** (Shao et al., 2009) | C/D: 118/132 | H/ACA: 43/69 |
| **Leishmania** (Liang et al., 2007) | C/D: 41/62 | H/ACA *A-like*: 0/37 |

# Discussion: statistics

- We presented a refinement of the training phase of snoReport using:
  - Different features from the candidate snoRNA sequence
  - More data from different vertebrate organisms
  - New versions of the tools and databases used to build the previous version of snoReport
- H/ACA box snoRNA classifier, with the same data training, had an improvement of:
  - 25% regarding to sensitivity
  - 13.85% regarding to F-score
- C/D box snoRNA classifier, with the same data training, had an improvement of:
  - 13.33% of sensitivity
  - 6.57% of F-score

# Discussion: validation

- We predicted 69.43% of all the sequences from different organisms
- Many sequences used in validation was not yet experimentally validated:
  - Maybe some of them are false positives, or
  - Are not representatives of the canonical snoRNAs (like the H/ACA box snoRNA of leishmania)
- This new version of snoReport is more efficient to identify both classes of snoRNAs, and can be used for many different organisms with high quality of prediction