X CAB2C

X Argentinian Conference on Bioinformatics and Computational Biology
November 13, 2019

MicroRNA prediction from genome-wide data with deep learning: a novel approach based on convolutional residual networks

C. Yones, L.A. Bugnon, J. Raad, D.H. Milone and G. Stegmayer

Research Institute for Signals, Systems and Computational Intelligence

UNL CONICET

sinc(*i*)

1. Introduction
   - Background
   - Motivation
2. Proposed method
   - Convolutional Neural Networks
   - Proposed architecture
3. Results
   - Experimental setup
   - Precision-recall curves
4. Conclusions

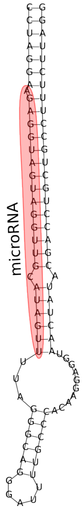
1 Introduction
- Background
- Motivation

2 Proposed method
- Convolutional Neural Networks
- Proposed architecture

3 Results
- Experimental setup
- Precision-recall curves

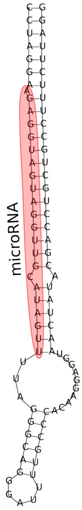
4 Conclusions

Background



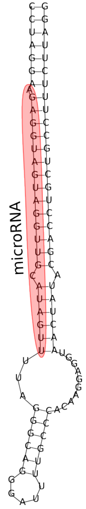
- MicroRNAs (miRNAs) play an essential role in post-transcriptional gene regulation.
- Precursors of miRNA (pre-miRNAs) are characterized by hairpins structure.

Background



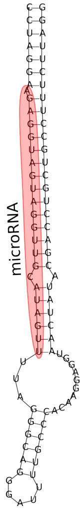
- MicroRNAs (miRNAs) play an essential role in post-transcriptional gene regulation.
- Precursors of miRNA (pre-miRNAs) are characterized by hairpins structure.
- A large amount of similar sequences can be folded into this kind of structure.

Background



- MicroRNAs (miRNAs) play an essential role in post-transcriptional gene regulation.
- Precursors of miRNA (pre-miRNAs) are characterized by hairpins structure.
- A large amount of similar sequences can be folded into this kind of structure.
- Machine learning algorithms have been proposed to predict which sequences are likely to contain a miRNA.

Background



- MicroRNAs (miRNAs) play an essential role in post-transcriptional gene regulation.
- Precursors of miRNA (pre-miRNAs) are characterized by hairpins structure.
- A large amount of similar sequences can be folded into this kind of structure.
- Machine learning algorithms have been proposed to predict which sequences are likely to contain a miRNA.

But, there are some problems with ML methods

- Datasets used are not representative of the wide variety of negative examples.
  $\rightarrow$ Use all hairpins of the genome for validation.

- The performance measures used underestimate the effect of imbalance.
  $\rightarrow$ Take into account the number of false positives.

But, there are some problems with ML methods

- Datasets used are not representative of the wide variety of negative examples.
  $\rightarrow$ Use all hairpins of the genome for validation.

- The performance measures used underestimate the effect of imbalance.
  $\rightarrow$ Take into account the number of false positives.

- The validation methodology does not imitate a real prediction task.
  $\rightarrow$ Test on sequences from unseen species.

But, there are some problems with ML methods

- Datasets used are not representative of the wide variety of negative examples.
  $\rightarrow$ Use all hairpins of the genome for validation.

- The performance measures used underestimate the effect of imbalance.
  $\rightarrow$ Take into account the number of false positives.

- The validation methodology does not imitate a real prediction task.
  $\rightarrow$ Test on sequences from unseen species.

Taking into account these points, simple sequence alignment works better than machine learning methods.

But, there are some problems with ML methods

- Datasets used are not representative of the wide variety of negative examples.
  $\rightarrow$ Use all hairpins of the genome for validation.

- The performance measures used underestimate the effect of imbalance.
  $\rightarrow$ Take into account the number of false positives.

- The validation methodology does not imitate a real prediction task.
  $\rightarrow$ Test on sequences from unseen species.

Taking into account these points, simple sequence alignment works better than machine learning methods.

Can it be done better?

> cel-mir-62

GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAG

(((((((((((..((((((.(((((.....)))))..))))))..)))))))))))..

Can it be done better?

> cel-mir-62

GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAG

(((((((((((..((((((.(((((.....)))))..))))))..)))))))))))..

> Unknown sequence 1

UUGAGCUAGAUUACAUAUCCGUAAGUGGGGCUUGAAUGAUAUGUAAUCUAGCUUACAG

.((((((((((((((((((..(((((...)))))...))))))))))))))))))...

Can it be done better?

> cel-mir-62

GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAG

(((((((((((..((((((.(((((.....)))))..))))))..)))))))))))..

> Unknown sequence 1

UUGAGCUAGAUUACAUAUCCGUAAGUGGGGCUUGAAUGAUAUGUAAUCUAGCUUACAG

.((((((((((((((((((..(((((...)))))...))))))))))))))))))...

> Unknown sequence 2

GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUCCGAUCUAAUCUACCUUACAG

(((((((((((..((.(((.............))).))..)))))).....)))))..

Can it be done better?

> cel-mir-62

GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAG

(((((((((((..((((((.(((((.....)))))..))))))..)))))))))))..

> Unknown sequence 1

UUGAGCUAGAUUACAUAUCCGUAAGUGGGGCUUGAAUGAUAUGUAAUCUAGCUUACAG

.((((((((((((((((((..(((((...)))))...))))))))))))))))))...

Matching score: 189 (clustalw2)

> Unknown sequence 2

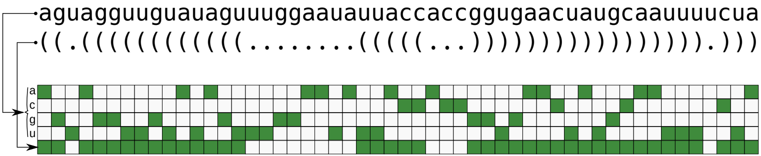
GUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUCCGAUCUAAUCUACCUUACAG

((((((((((((..((.(((...........))).))..))))))).....)))))..

Matching score: 249 (clustalw2)

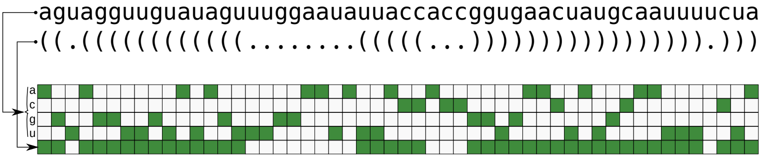
1. Introduction
   - Background
   - Motivation
2. Proposed method
   - Convolutional Neural Networks
   - Proposed architecture
3. Results
   - Experimental setup
   - Precision-recall curves
4. Conclusions

A gentle introduction to Convolutional Neural Networks



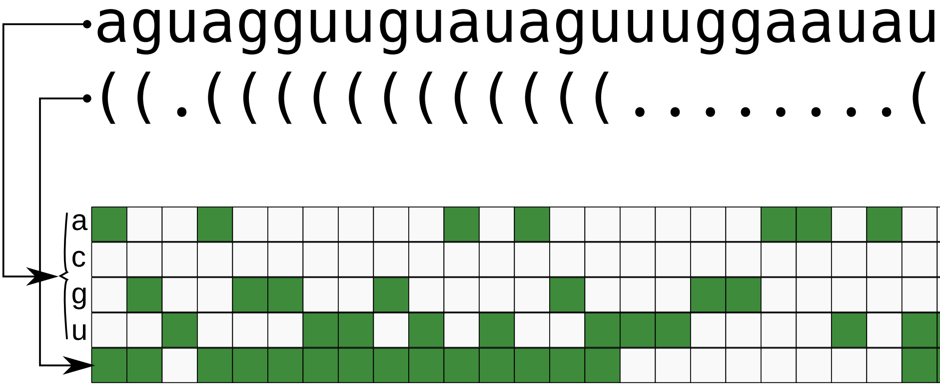
Each column represents a nucleotide

A gentle introduction to Convolutional Neural Networks

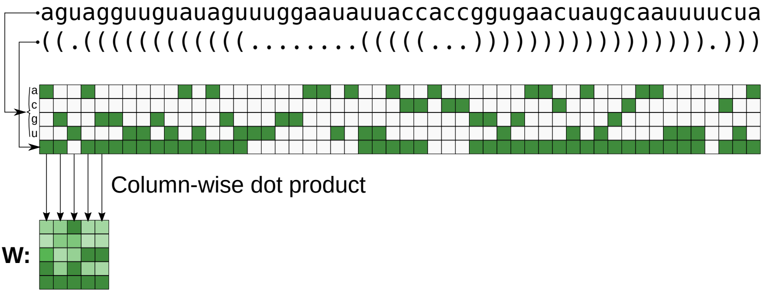


Each column represents a nucleotide

A gentle introduction to Convolutional Neural Networks

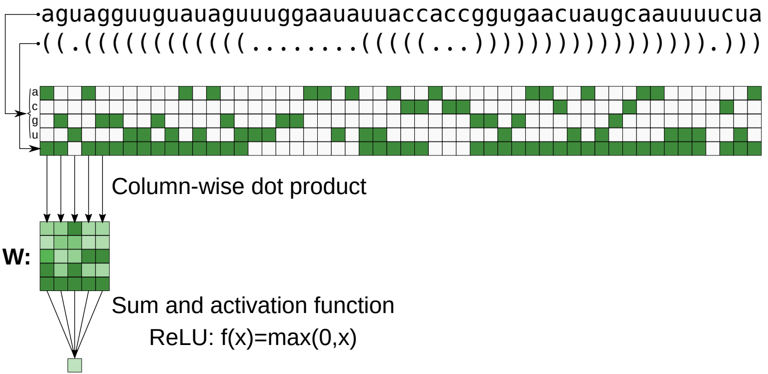


A gentle introduction to Convolutional Neural Networks



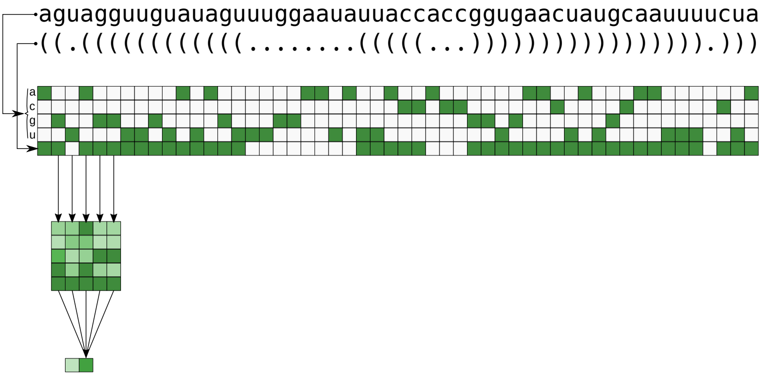
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



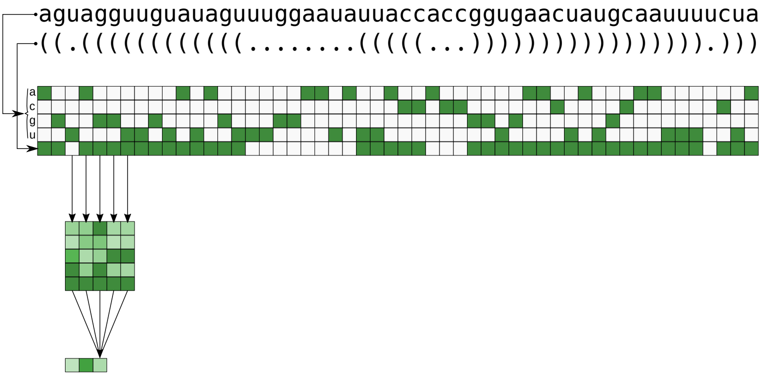
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



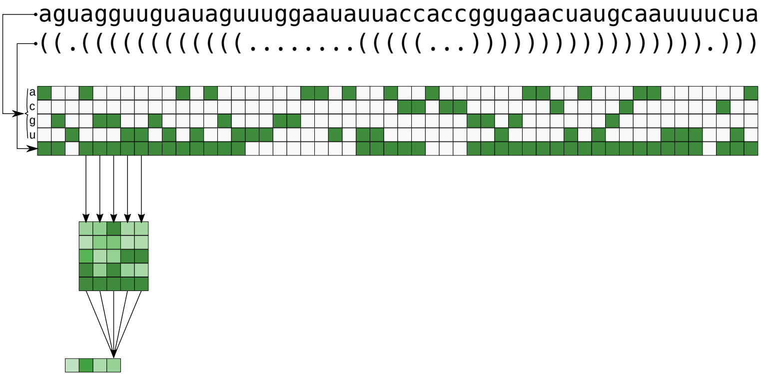
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



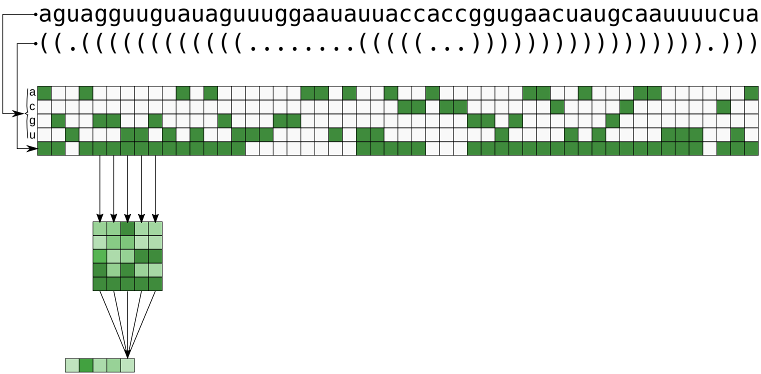
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



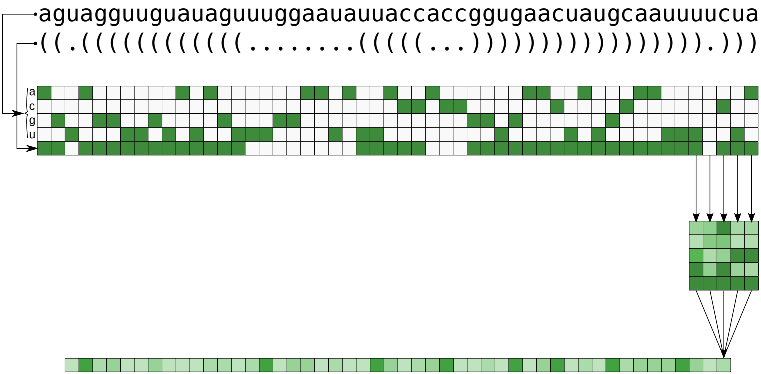
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



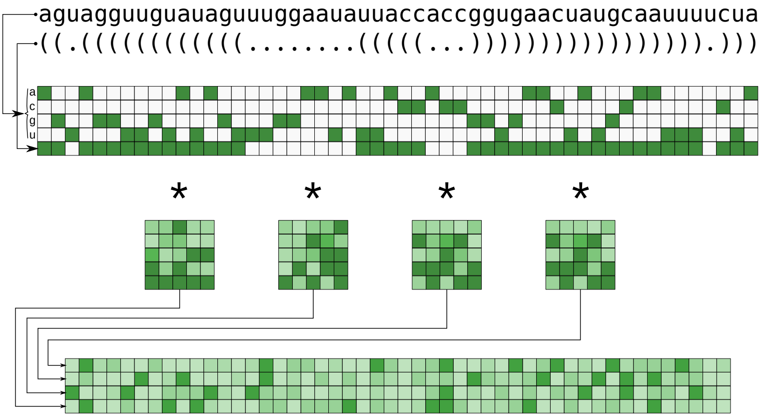
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



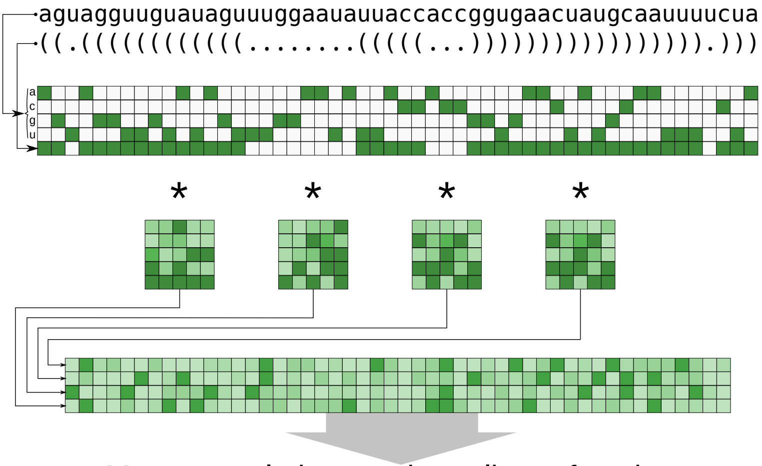
Convolution of the input tensor with the filter **W**

A gentle introduction to Convolutional Neural Networks



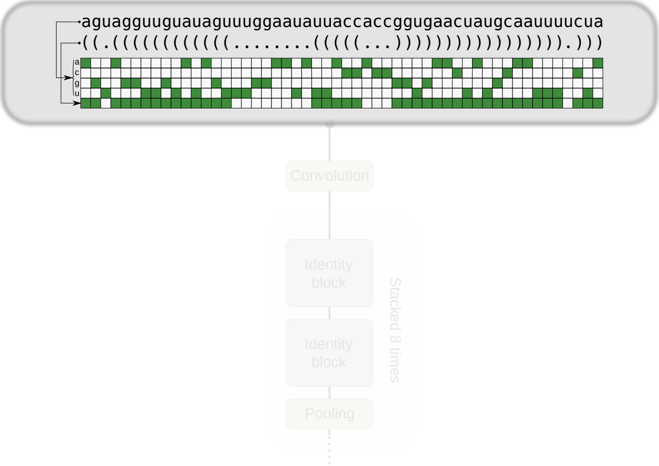
Each filter generates a row in the output tensor

A gentle introduction to Convolutional Neural Networks

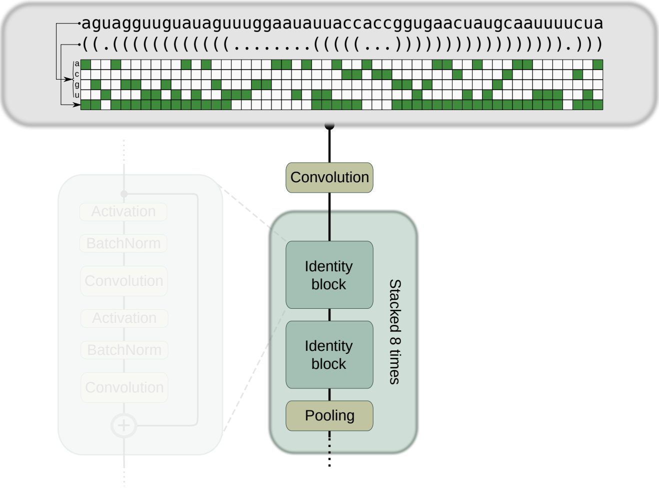


More convolutions and non-linear functions

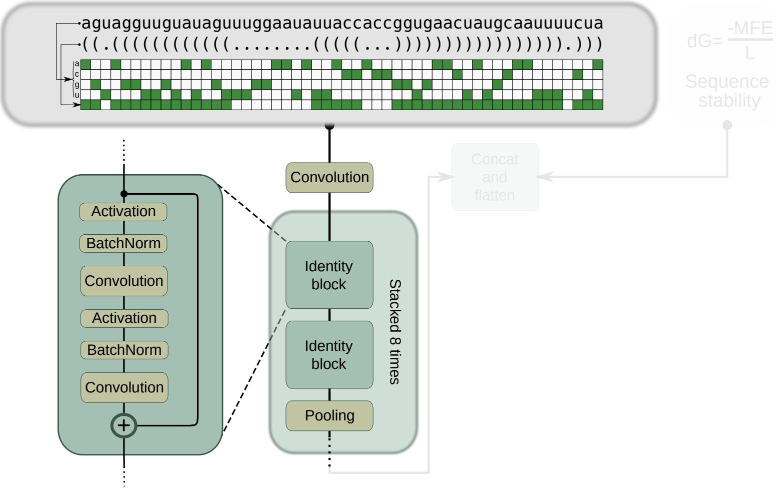
Proposed architecture: mirDNN



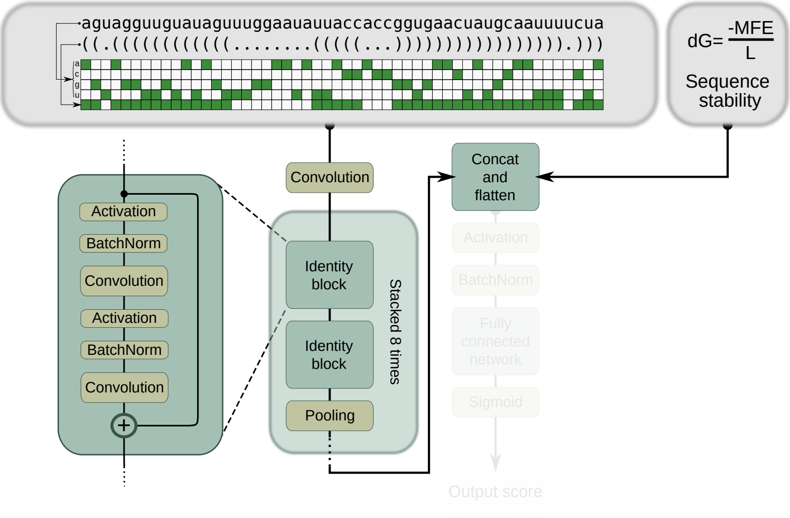
Proposed architecture: mirDNN



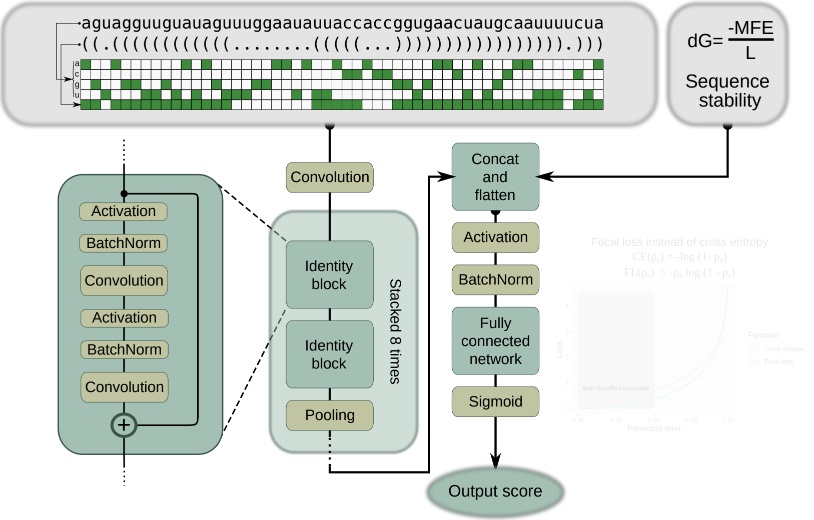
Proposed architecture: mirDNN



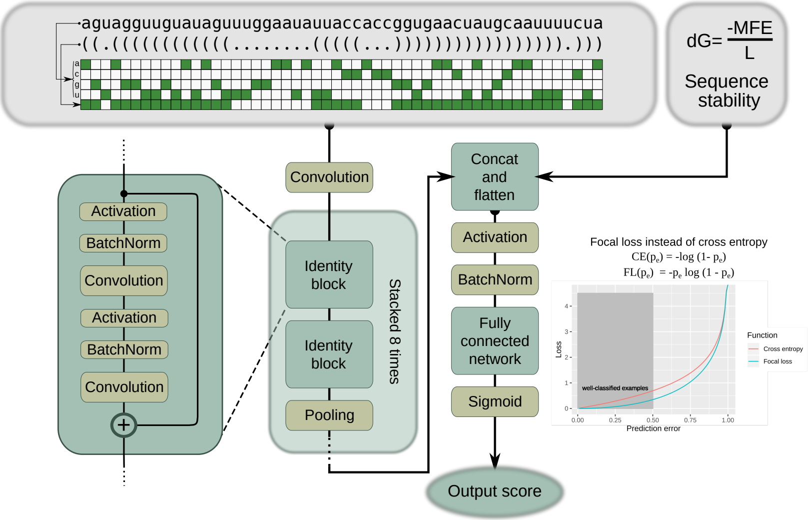
Proposed architecture: mirDNN



Proposed architecture: mirDNN



Proposed architecture: mirDNN



Proposed architecture: mirDNN

on

rm

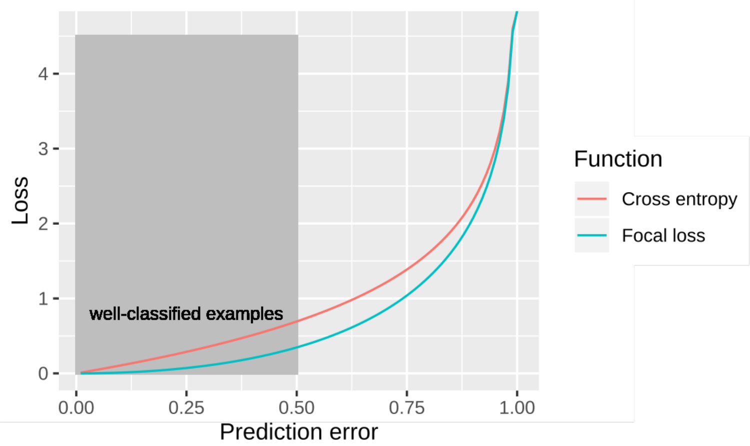
ed
k

d

Focal loss instead of cross entropy

$$CE(p_e) = -\log(1 - p_e)$$

$$FL(p_e) = -p_e \log(1 - p_e)$$



1 Introduction
- Background
- Motivation

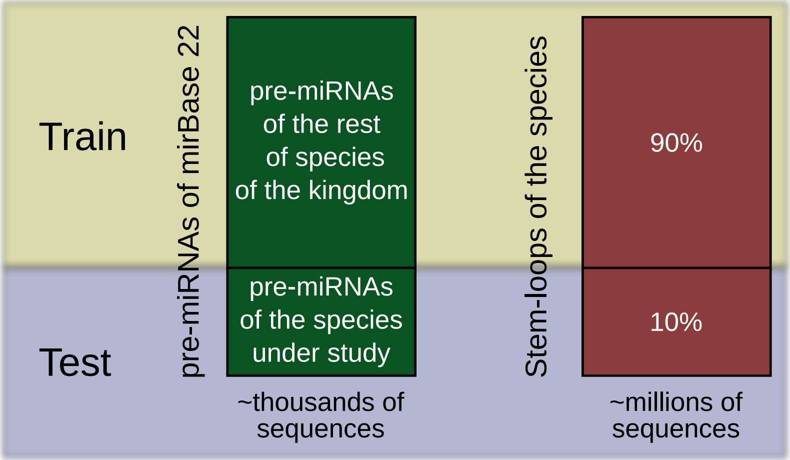
2 Proposed method
- Convolutional Neural Networks
- Proposed architecture

3 **Results**
- Experimental setup
- Precision-recall curves

4 Conclusions

Experimental setup

Validation on a *leave-species-out* scheme



Experimental setup

- Test on three well-known species: *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Anopheles gambiae*.
- Two machine learning methods and one sequence alignment method were used for comparison.

Experimental setup

- Test on three well-known species: *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Anopheles gambiae*.
- Two machine learning methods and one sequence alignment method were used for comparison.
- The precision and recall were used as performance measures:

$$Pr = \frac{TP}{TP+FP} \qquad Rc = \frac{TP}{TP+FN}$$

Experimental setup

- Test on three well-known species: *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Anopheles gambiae*.
- Two machine learning methods and one sequence alignment method were used for comparison.
- The precision and recall were used as performance measures:

$$Pr = \frac{TP}{TP+FP} \qquad Rc = \frac{TP}{TP+FN}$$

- Varying the threshold that defines what is classified as positive or negative, precision-recall curves were generated.

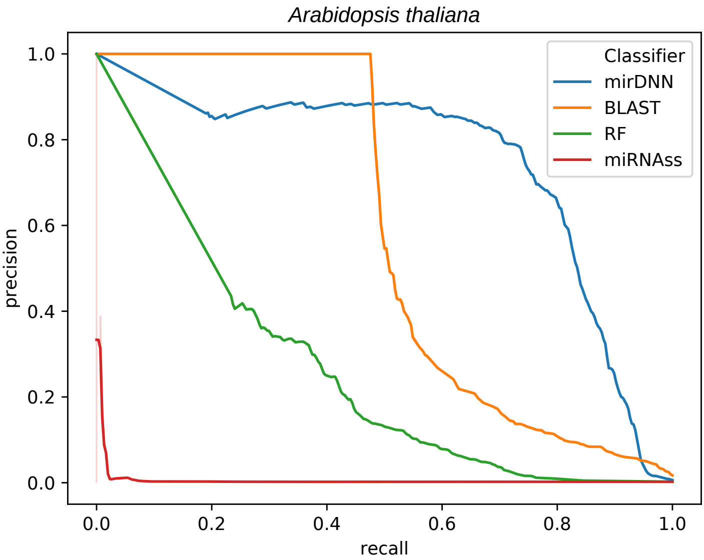
Experimental setup

- Test on three well-known species: *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Anopheles gambiae*.
- Two machine learning methods and one sequence alignment method were used for comparison.
- The precision and recall were used as performance measures:

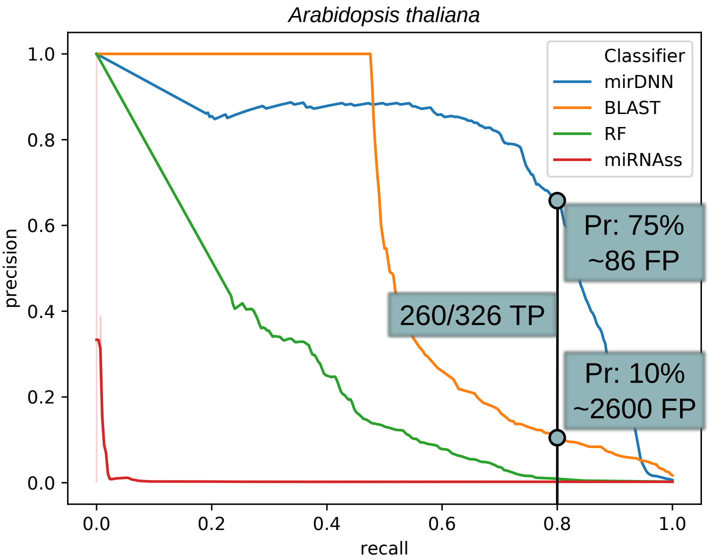
$$Pr = \frac{TP}{TP+FP} \qquad Rc = \frac{TP}{TP+FN}$$

- Varying the threshold that defines what is classified as positive or negative, precision-recall curves were generated.

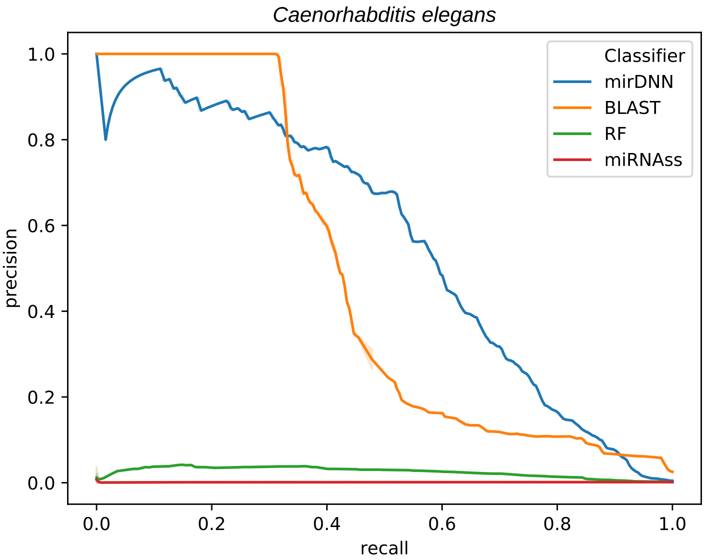
Precision-recall curves



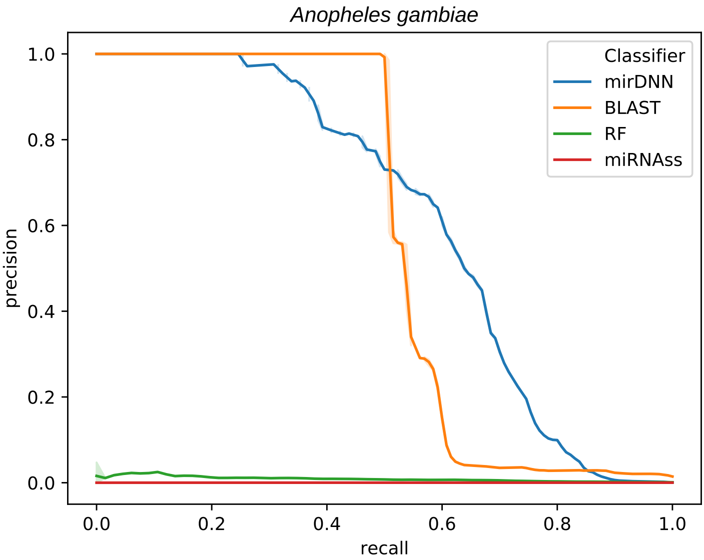
Precision-recall curves



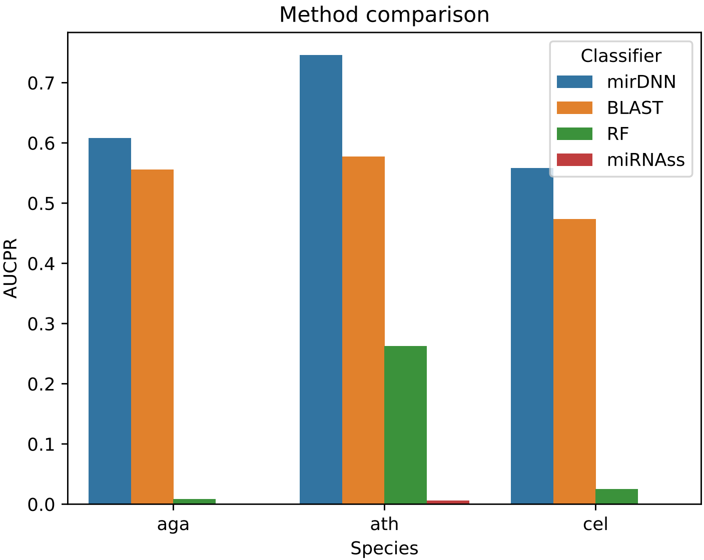
Precision-recall curves



Precision-recall curves



Area under the curves



1. Introduction
   - Background
   - Motivation
2. Proposed method
   - Convolutional Neural Networks
   - Proposed architecture
3. Results
   - Experimental setup
   - Precision-recall curves
4. Conclusions

Conclusions

- State-of-the-art machine learning methods for miRNA prediction do not perform well on new species.
- Sequence alignment methods have a better performance, but only for low recall rates.

Conclusions

- State-of-the-art machine learning methods for miRNA prediction do not perform well on new species.
- Sequence alignment methods have a better performance, but only for low recall rates.
- Using convolutional neural networks, we can *learn* which are the important characteristics that define a pre-miRNA.

Conclusions

- State-of-the-art machine learning methods for miRNA prediction do not perform well on new species.
- Sequence alignment methods have a better performance, but only for low recall rates.
- Using convolutional neural networks, we can *learn* which are the important characteristics that define a pre-miRNA.
- The proposed method achieves a precision many times higher at high recall rates in three tested species.

Conclusions

- State-of-the-art machine learning methods for miRNA prediction do not perform well on new species.
- Sequence alignment methods have a better performance, but only for low recall rates.
- Using convolutional neural networks, we can *learn* which are the important characteristics that define a pre-miRNA.
- The proposed method achieves a precision many times higher at high recall rates in three tested species.

X CAB2C

X Argentinian Conference on Bioinformatics and Computational Biology
November 13, 2019

MicroRNA prediction from genome-wide data with deep learning: a novel approach based on convolutional residual networks

C. Yones, L.A. Bugnon, J. Raad, D.H. Milone and G. Stegmayer

Research Institute for Signals, Systems and Computational Intelligence

UNL
CONICET
sinc(i)

cyones@sinc.unl.edu.ar
http://www.sinc.unl.edu.ar/