One-class SVM to identify candidates to Rerefence Genes based on the augment of RNA-seq data with Generative Adversarial Networks

Edwin J. Rueda Rommel Ramos

Edian F. Franco Orlando Belo Jefferson Morais

Edwin J. Rueda
Federal University of Para
edwin.rojas@icen.ufpa.br

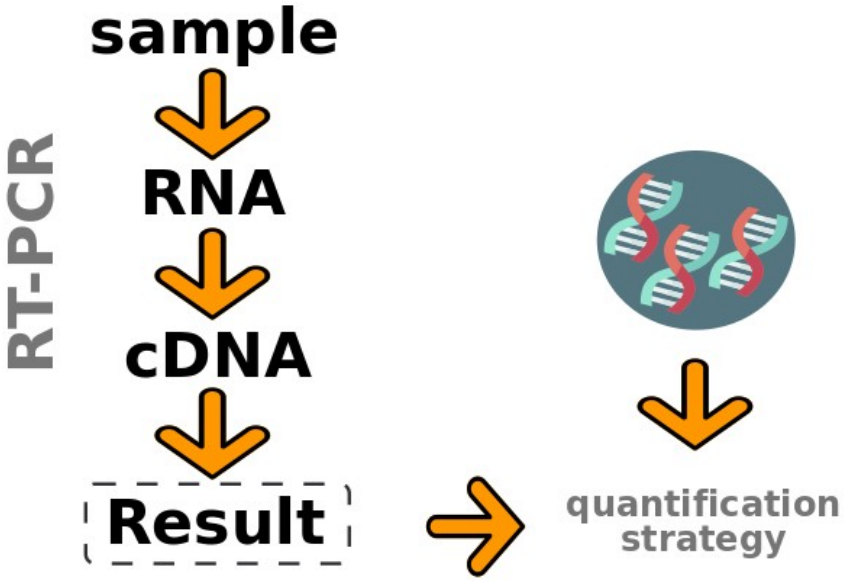
July 1-4, 2020

SUMMARY

- Introduction
- Proposed Method
- Experiments and Results
- Conclusions and Future Works

Introduction

- **Reference Genes (RG)**
  - Constitutive genes
  - Expressed in all cells
  - Used in internal controls (gene expression analysis)

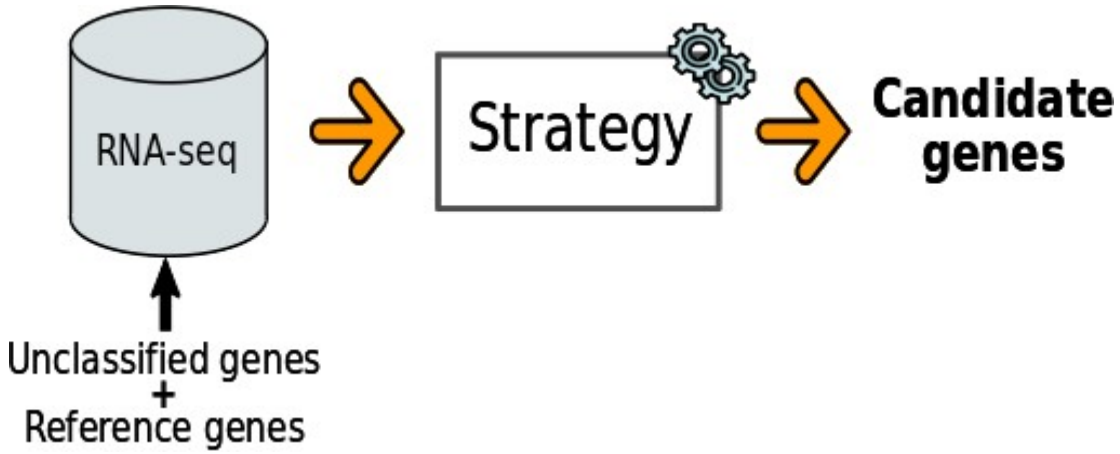


- Normalize gene expression
- Demonstrates the variability and imperfections of the technology

Introduction

Identify candidates for Reference Genes

- Pipeline:

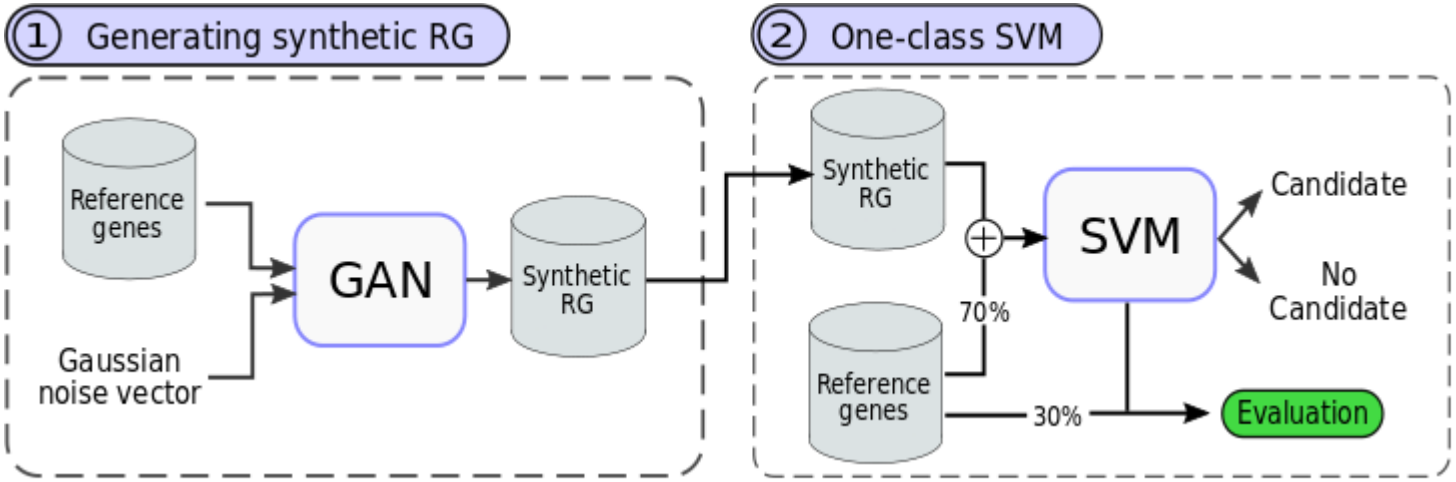


- Strategies:
  - Based on clustering (Euclidean distance)
  - Based on optimization algorithms (Euclidean distance*)

Proposed Method

- Steps:

  1) Generative Adversarial Networks to augment of RNA-seq data
  2) One-Class SVM to select candidate genes



- Initial step: data processing

Proposed Method

Data processing

- Pipeline:

1) Normalization with RPKM:

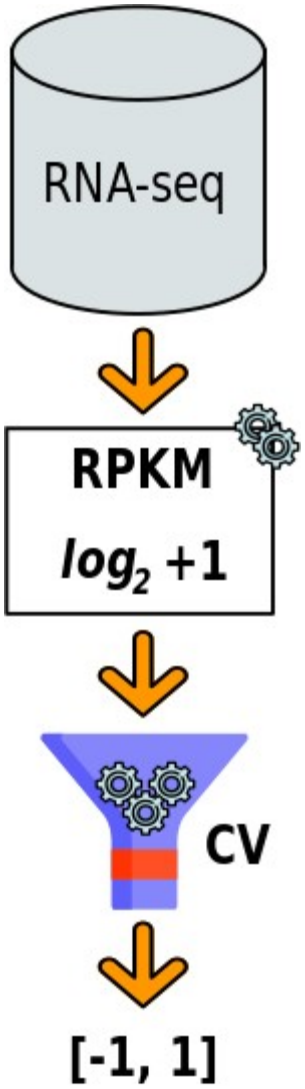
$$RPKM = \frac{numReads * 10^9}{geneLength * TMReads}$$

2) $\log_2 +1$

3) Remove outliers based on the Coefficient of Variation (CV):

$$CV = \frac{\sigma}{\mu}$$

4) Data scaled between [-1, 1]

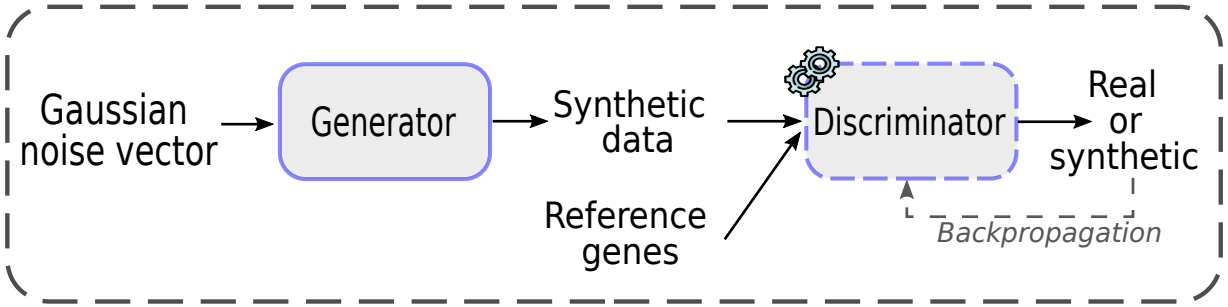


Proposed Method

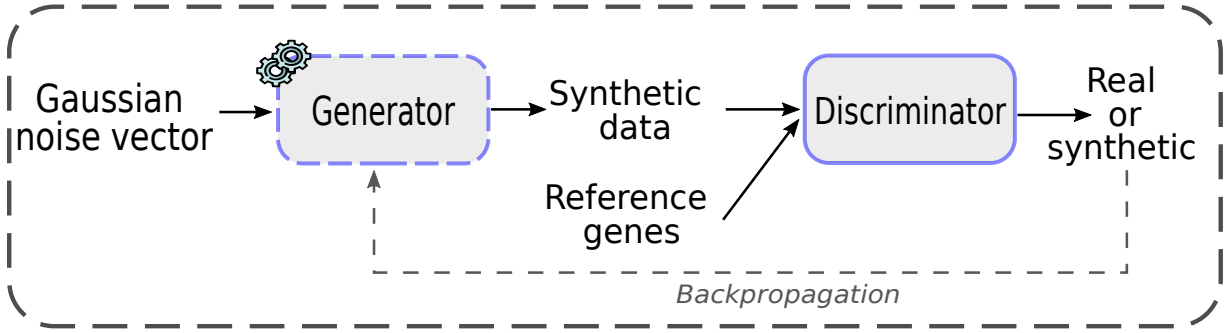
Generative Adversarial Networks (GAN)

- Steps:

1) Training the Discriminator network and freezing their weights



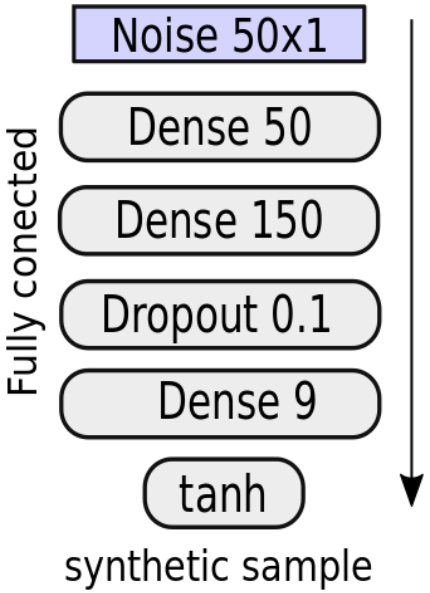
2) Training the Generator network



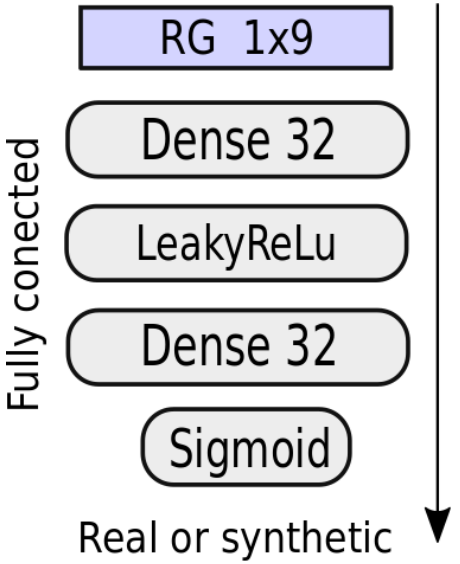
Proposed Method

Proposed architecture

- Generator network



- Discriminator network



- A normal distribution *N(0,1)* as a noise vector
- Stochastic Gradient Descent to compute the gradients

Proposed Method

Evaluation

- A proposed Similarity metric ***S(x,x')*** to evaluate the performance of the GAN:

$$S(x, x') = \sum_{i}^{m} \sum_{j}^{n_g} \sum_{k}^{n_f} \frac{|x_i^{(k)} - x'^{(k)}_j|}{n_f n_g m} + |0.5 - \frac{1}{n_g} \sum_{j}^{n_g} \hat{y}_j|$$

- $\hat{y}$: Class predicted by *D* network for a synthetic gene
- x' : Set of synthetic genes generated by the *G* network
- x : Set of Reference Genes
- m : Number of Reference Genes
- n_g : Number of synthetic genes
- n_f : Number of features (gene expression)

Proposed Method

Evaluation

- A proposed ***E(x')*** metric to select the best sample of synthetic data:

$$E(x') = \frac{1}{n_g} \sum_{j}^{n_g} \left[CV(x'_j) + \frac{1 - D(x'_j)}{D(x'_j)} \right]$$

- CV : Coefficient of variation
- x' : Set of synthetic genes generated by the *G* network
- $n_g = 300$: Number of synthetic genes
- **Other metrics:** Binary Cross-Entropy, Precision score

Proposed Method

One-Class SVM

- Implements the RBF kernel (Gaussian kernel)
- Recall score to evaluate the performance of the One-Class SVM:

$$Recall = \frac{TP}{TP + FN}$$

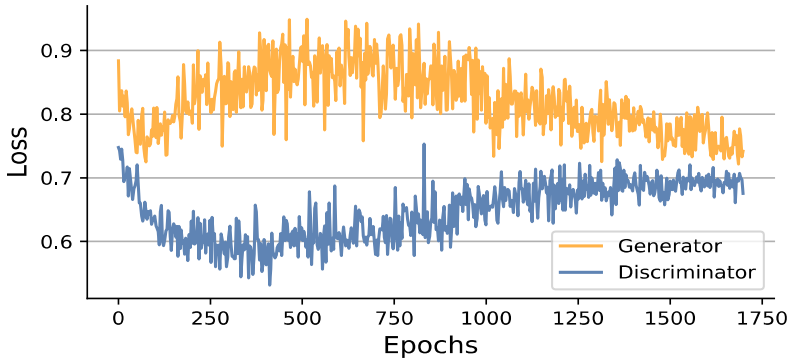
- TP : True Positives
- FN : False Negatives
- Recall score allows to measure the ability of the classifier to find all positive samples (RG)
- A Recall score close to one indicates that the classifier has a good performance

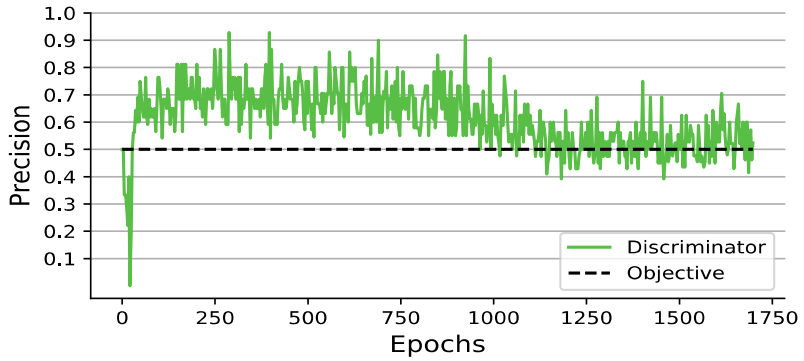
Experiments and Results

- This approach was evaluated with the *Escherichia coli* MG1655 dataset
- Parameters used for training the GAN architecture

Parameter	Generator	Discriminator
Optimizer	SGD	SGD
Learning rate	0.00015	0.001
Decay rate	0.00015/1700	0.001/1700
Momentum	0.92	0.9
Epochs	1700	1700

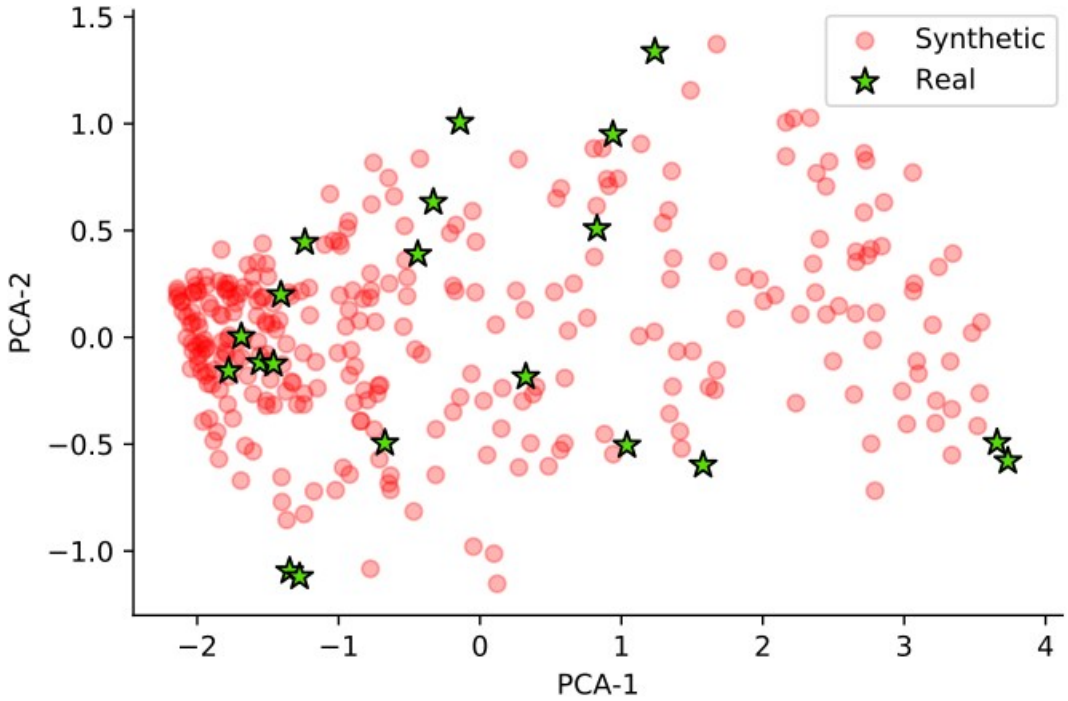
- Convergence process





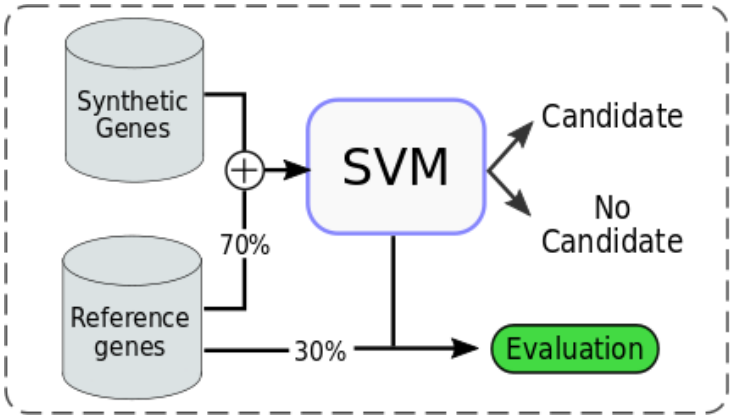
Experiments and Results

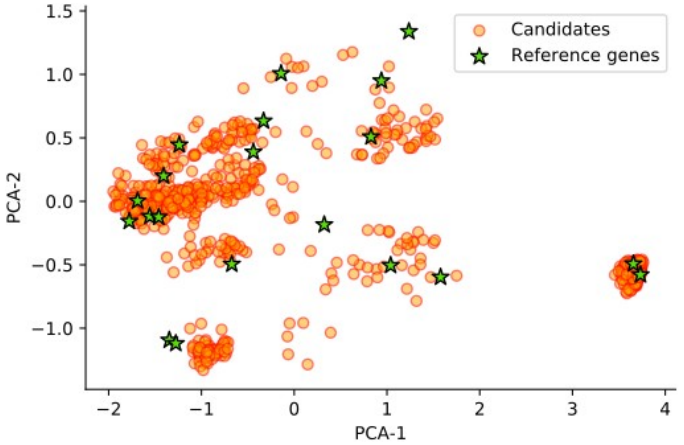
- Selection of synthetic genes
  - We generated 5000 sets of 300 synthetic genes and selected the set with the smallest value of ***E(x')***



Experiments and Results

- Selection of candidate genes





- *Recall* score on test and training data

Metrics	Training data		Test data	
	Augmented data	Reference Genes (70%)	Reference Genes (30%)	Synthetic data
Recall	98.40%	92.85%	85.71%	98.26%

- We selected the 11 most relevant candidate genes

Experiments and Results

- Comparing results with augmented and unaugmented data:

Reference Genes	*Recall score*	
	Training data	Test data
Augmented	98.40%	85.71%
Unaugmented	85.71%	66.66%

- With augmented dataset, we improved recall score of proposed classifier by 19%

Conclusions

- With the proposed method we were able to identify 807 possible candidate genes from a total of 4170 unclassified genes from the *Escherichia coli* MG1655 dataset.
- Augmenting the set of RG we can increase the performance of the One-class svm classifier
- Code is being maintained on GitHub: https://github.com/ejrueda/20_ICCSA_RG_GAN

Future Works

- Other GAN architecture can be tested
- A more extensive hyperparameter tuning could also be applied in the training of the GAN architecture
- Other One-class classifiers can be tested (Ex: GAN)

One-class SVM to identify candidates to Rerefence Genes based on the augment of RNA-seq data with Generative Adversarial Networks

Thank you!

Edwin J. Rueda
Federal University of Para
edwin.rojas@icen.ufpa.br

July 1-4, 2020