Evaluation of Inductive and Transductive Inference in the context of Translation Initiation Site

Wallison W. Guimarães, Cristiano L. N. Pinto, Cristiane N. Nobre, Luis E. Zárate
Pontifical Catholic University of Minas Gerais, Brazil
wallison.guimaraes@sga.pucminas.br,cristiano@emge.edu.br,{nobre,zarate}@pucminas.br

ABSTRACT

The prediction of Translation Initiation Site (TIS) from a mRNA (Ribonucleic Acid Messenger) is a relevant and latent problem of molecular biology, which has benefited from the evolution of computational techniques of machine learning (ML). There are some machine learning scenarios where the dataset either does not have enough classified sequences to train a precise model, or it does not have an *upstream* region, such as *Caenorhabditis elegans*. In this article, we compare the inductive and transductive approaches for TIS prediction, using a methodology that disregards the *upstream* region. With the proposed methodology, we achieved 95% training accuracy, using only 2.5% of sequences belonging to the *Caenorhabditis elegans* class, which has many available sequences but does not have the *upstream* region, and 75% for the *Rattus norvegicus* class, which has fewer sequences available, using a transductive approach. Our results demonstrate the viability of the transductive approach for scenarios with fewer sequences, a common situation for organisms with incomplete gene sequencing.

CCS CONCEPTS

• **Theory of computation → Semi-supervised learning**; *Support vector machines*; • **Applied computing → Bioinformatics**; • **Computing methodologies** → Supervised learning by classification;

KEYWORDS

Transductive Inference, Prediction of Translation Initiation Sites, Support Vector Machine, Inductive inference, Upstream and Downstream regions

ACM Reference format:
Wallison W. Guimarães, Cristiano L. N. Pinto, Cristiane N. Nobre, Luis E. Zárate. 2018. Evaluation of Inductive and Transductive Inference in the context of Translation Initiation Site. In *Proceedings of ACM SAC Conference, Pau,France, April 9-13, 2018 (SAC'18),* 4 pages.
https://doi.org/10.1145/3167132.3167368

1 INTRODUCTION

The prediction of the Translation Initiation Site (TIS) from a mRNA (Ribonucleic Acid Messenger) is a relevant and latent problem of molecular biology, which has benefited from the evolution machine learning (ML) computational techniques.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SAC'18, April 9-13, 2018, Pau,France
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5191-1/18/04…$15.00
https://doi.org/10.1145/3167132.3167368

The translation and transcription processes are the way cells transmit and express their genetic information [11]. However, only a few parts of the transcribed sequence carry the necessary information to encode the proteins. These sequences are called CoDing Sequence (CDS). The CoDing Sequence (CDS) region is delimited by flags called start codons, usually identified by the AUG triplet and also responsible for the start of the protein synthesis process, and stop codon, usually identified by the occurrence of the UAA, UAG or UGA triplets [7]. The positions of the start and stop codons directly influence the structure and function of the proteins produced, which makes the correct identification of the TIS considered a central problem in computational biology.

The concepts of *upstream* region, *downstream* region, and the reading phase of the mRNA sequence by the ribosome were proposed by Kozak [1984]. The negative sequences are categorized as *upstream* in phase (UPIP) or CDS in phase (CDSIP), when these are read in the same phase of the TIS and *upstream* out of phase (UPOP) or CDS out of phase (CDSOP) otherwise.

The Kozak consensus [6] revealed that the -3 and +4 positions of the mRNA are conservative, having a predominance of nucleotides A/G and G, respectively. These positions are considered by several authors in the definition of the methodology of nucleotide window extraction for the prediction of TIS [9] [8].

In the methodology proposed by Pinto et al. [2017], the results achieved by their model exceed the results achieved by the *TISHunter*[1], *TIS Miner*[2], and *NetStart*[3] tools, broadly used in TIS prediction applications. The authors define the beginning of the windowing from nucleotide -9, that is, 9 positions in the *upstream* region, ensuring the presence of the Kozak consensus [6] on the the extracted window. This methodology, to a certain extent, limits the number of extracted sequences, as well as the number of organisms analysed, since it discards molecules that do not contain 9 (nine) nucleotides in the TIS *upstream* region. In other words, organisms that do not have a sequenced *upstream* region can't be analysed using their methodology. This is the case, for example, of the organism *Caenorhabditis elegans*, which has only the *downstream* region of the molecule.

The process of computationally identifying the AUG (start) codon depends on a method capable of predicting or classifying the positive class (TIS) and negative class examples (codon not indicative of Translation Initiation Site - nTIS). A classification algorithm that has been frequently used in TIS prediction problems is the Support Vector Machine (SVM), which can be based on both inductive inference and transductive inference [8].

[1] Available at http://tishunter.ucr.edu/
[2] Available at http://dnafsminer.bic.nus.edu.sg/Tis.html
[3] Available at http://www.cbs.dtu.dk/services/NetStart/

Organisms that have few sequenced molecules are a challenge for inductive models because of the lack of information, and the semi-supervised transductive inference [2] is ideal for such scenarios. The main objective of transductive inference is to construct a classifier using two sets of data simultaneously: the original training set, with already classified (labelled) data, and the prediction dataset (unlabelled), with records that do not have assigned labels yet. By using both datasets at the same time, the transductive inference has more training information available than inductive inference and can be considered as an alternative approach to tackle the TIS prediction problem.

The objective of this paper is to evaluate the inductive and transductive approaches in the prediction context of TIS. For this, we propose a methodology that, unlike the one used in Pinto et al. [2017], disregards the *upstream* region. This strategy allows for the analysis of organisms that do not have a sequenced *upstream* region, such as the *Caenorhabditis elegans*. In addition to that, out methodology considers the nTIS sequences of the *downstream* region, which is also different from the work of Pinto et al. [2017], which considered nTIS of the *upstream* region.

This article is organized as follows: Section 2 presents related works. Section 3 we provide a discussion of the materials and methods used. The experimental results are shown in Section 4. Finally, Section 5 contains our conclusions and future work.

2 RELATED WORK

The transductive inference has been applied across a wide variety of domains, including feelings analysis, text classification, referral systems, natural language processing, image processing, and the prediction or diagnosis of various events in medical fields [5], [10]. In the area of Bioinformatics, transductive approaches have been successfully used mainly for problems related to proteins.

In Pinto et al. [2017], the authors have applied the use of transductive inference to predict TIS. The authors concluded that the transductive inference approach, applied in scenarios with few available training sequences, is feasible and achieves superior accuracy to inductive inference (ISVM). The authors tested different extraction window sizes, and the most accurate extraction window had 1081 nucleotides in the *downstream* region. However, their results limit the applicability of their methodology exclusively to organisms that have the *upstream* region.

Stanescu and Caragea [2015] did a comparative study of transductive algorithms for prediction of splicing sites of the organism *Caenorhabditis elegans*. The authors compared the Transductive Support Vector Machines (TSVM) [2], Label Propagation and a variant of the Adsorption algorithm, the Modified Adsorption (MAD). They evaluated the representativeness of the 5%, 10%, 15% and 20% size of the labelled dataset in the training dataset, using 5-Fold Cross-Validation. Their results showed that the TSVM outperformed the other algorithms when trained with less labelled data, more precisely when the labelled data represented between 5% to 15% of the training base.

In Kondratovich et al. [2013] the authors use TSVM in a prediction problem of molecule activity. The authors use small and unbalanced datasets, reaching results that showed that the TSVM is efficient in this type of scenario.

3 MATERIALS AND METHODS

3.1 Dataset

The datasets used in this paper correspond to the datasets used in Pinto et al. [2017]. The data was extracted from the public database *RefSeq*, from NCBI[4] on April 22, 2014. The records on the dataset have the level of inspection *reviewed*[5] and refer to the organisms *Rattus norvegicus* (167 molecules), *Mus musculus* (1.025 molecules), *Homo sapiens* (25.801 molecules), *Drosophila melanogaster* (35.847 molecules) and *Arabidopsis thaliana* (22.173 molecules). In addition to these organisms, we also consider the *Caenorhabditis elegans* organism (26.066 molecules) which only has the *downstream* region (meaning all the mRNA sequences available for this organism started with the TIS sequence).

For the input of our model, we consider the AUGs located at TIS position as positive class and the AUGs located at CDSOP region as negative class (nTIS). Preliminary tests have shown that using CDSOP sequences as nTIS, rather than CDSIP, UPOP or UPIP, provides an improvement in evaluation metrics. Starting at each AUGs (TIS and nTIS) we extract a window of 1081 nucleotides (*downstream*).

This window of 1081 nucleotides was considered by de Pinto et al. [2017] and got the best results in their tests. All sequences were encoded in binary chains, where each nucleotide was encoded as a 4-bit chain (A = 1000, C = 0100, G = 0010 and U = 0001), respectively. This coding is also used in [3], [9], and [8].

The TIS context prediction induces a natural class imbalance, since for each mRNA molecule there is only one AUG codon identified as the start codon (TIS), while all other AUG codons are identified as nTIS. The class imbalances of *M. musculus* and *R. norvegicus*, for example, are in the order of 1:23 and 1:131, respectively.

The strategy adopted to work around the class imbalance problem of the TIS and nTIS sequences was the random undersampling method, which got good results and did not provide computational performance problems for the base balancing.

3.2 TSVM and Parameter Definitions

Semi-supervised learning techniques, as mentioned, make use of unlabelled data during training. Generally, this strategy is used in contexts where there is a small amount of labelled data, such as the TIS prediction problem, where the unlabelled data corresponds to the new sequenced molecules whose TIS have not yet been identified. It is important to note that the TIS identification process usually requires the participation of a human expert or biochemical experiments, which makes the labelling process expensive and complex. This reinforces the need for a reliable tool that automates TIS identification, such as the use of Transductive SVM (TSVM).

According to Vapnik et al. [1998], the definition of "transduction" is the inference of the particular to the particular. For example, in a classification problem, this means that given the classifications y_i, $i = 1, ..., l$, and l labelled examples of the training set $x_i, ..., x_l$, the Transductive model should predict the classification of k unlabelled examples $x_{l+1}, ..., x_{l+k}$ from a prediction set, during the learning process. In an Inductive inference model, the aim is to first

[4]Available at http://www.ncbi.nlm.nih.gov

[5]The description of status is available http://www.ncbi.nlm.nih.gov/books/NBK21091/

find a function that can describe the problem, and then classify the prediction set.

We utilized the ISVM and TSVM with the Gaussian kernel function RBF (Radial Basis Function), defined by Equation[6].

For the adjustment of the C and γ parameters[7], we used the Grid Search algorithm [8] [1] with 10% of the dataset.

3.3 Validation and evaluation metrics

To compare the performance of the ISVM and TSVM methods on a context that is favourable to the use of transductive models, the experiments are performed in two different scenarios. The traditional 10-fold cross-validation method [4] is referred throughout this work as "Scenario 1", and "Scenario 2" consists of an inversion of the standard cross-validation, that is, 1 fold is used as the training dataset, and the 9 remaining folds are used for testing.

We considered 3 evaluation metrics: *Precision*[9], *Sensitivity*[10], and *F-measure*[11].

4 RESULTS

The experiments began with the preprocessing of the datasets described in Section 3.1. This stage is summarized as the extraction and coding of sequences, according to the methodology proposed in this work. After that, the training and validation sets, for Scenarios 1 and 2, as defined in Section 3.3, were constructed for each evaluated organism. It is worthwhile to note that the data selected for both Scenarios 1 and 2 were submitted to the training of both the ISVM and TSVM.

The ISVM and TSVM classifiers were submitted to training for the organisms *Rattus norvegicus*, *Mus musculus*, *Homo sapiens*, *Drosophila melanogaster*, *Arabitopsis thaliana* and *Caenorhabditis elegans*. This last mentioned organism is characterized by not having an *upstream* region. The performance of the classifiers for each organism for the two scenarios was then evaluated and compared using the metrics described in Section 3.3.

The results obtained for each experiment are shown in Table 1. For Scenario 1, the results indicate that the performances of the ISVM and TSVM classifiers are very close when submitted to a training dataset where 90% of the sequences are available for training. But it can be observed that the accuracy of the TSVM was slightly higher than the accuracy for the ISVM model, reaching a little over 2% increase for the organism *Rattus norvegicus*, which has the lowest number of sequenced molecules.

For Scenario 2, where only 10% of the sequences are used as the training dataset, the performances of the two classifiers decrease, which was expected. The performance of the TSVM remains slightly higher than the ISVM. Analysing the results of Scenario 2 on Table 1, it can be seen that, similarly to the Scenario 1, the accuracy of the TSVM, was slightly higher in all organisms.

[6] $K(x_i, x_j) = exp^{-\frac{1}{2\sigma^2}||x_i - x_j||^2}$

[7] ***R. norvegicus*** and ***M. musculus***: Gamma=1.220703125 x 10^{-4} and C=8, ***H. sapiens***: Gamma=4.8828125 x 10^{-4} and C=8, ***C. elegans***: Gamma=4.8828125 x 10^{-4} and C=2, ***D. melanogaster***: Gamma=3.0517578125 x 10^{-5} and C=8, ***A. thaliana***: Gamma=3.0517578125 x 10^{-5} and C=32

[8] Available at https://www.csie.ntu.edu.tw/ cjlin/libsvm/

[9] $Pr = \frac{TP}{TP + FP}$

[10] $Se = \frac{TP}{TP + FN}$

[11] $F-measure = 2 \times \frac{Pr \times Se}{Pr + Se}$, where the TP, TN, FP, and FN are the numbers of True Positives, True Negatives, False Positives and False Negatives, respectively.

Table 1: Validation results for the scenarios 1 and 2

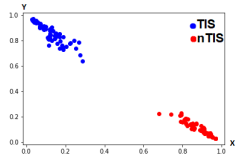
Scenario 1						
Organisms	**Inductive**			**Transductive**		
	Precision	Sensitivity	*F-measure*	Precision	Sensitivity	*F-measure*
R. norvegicus	91.53 ± 14.76	100.0 ± 0.00	94.86 ± 9.30	93.79 ± 13.99	98.33 ± 5.00	95.39 ± 9.65
M. musculus	97.83 ± 2.87	99.23 ± 1.18	98.49 ± 1.43	97.72 ± 1.76	98.21 ± 1.64	97.95 ± 1.54
H. sapiens	98.85 ± 0.44	99.42 ± 0.36	99.13 ± 0.35	98.99 ± 0.29	99.01 ± 0.29	99.00 ± 0.29
D. melanogaster	98.75 ± 0.32	99.24 ± 0.24	98.99 ± 0.23	99.04 ± 0.39	99.11 ± 0.28	99.07 ± 0.33
A. thaliana	99.66 ± 0.17	99.79 ± 0.09	99.72 ± 0.10	99.68 ± 0.13	99.69 ± 0.15	99.69 ± 0.14
C. elegans	97.81 ± 0.42	99.06 ± 0.28	98.43 ± 0.22	98.37 ± 0.29	98.39 ± 0.26	98.38 ± 0.27
Scenario 2						
Organisms	Precision	Sensitivity	F-measure	Precision	Sensitivity	F-measure
R. norvegicus	83.06 ± 10.91	80.82 ± 14.11	80.10 ± 7.80	83.19 ± 4.71	83.97 ± 2.88	83.55 ± 3.68
M. musculus	95.58 ± 1.46	98.72 ± 0.46	97.12 ± 0.65	97.11 ± 0.72	94.14 ± 0.65	97.12 ± 0.68
H. sapiens	96.44 ± 0.44	99.00 ± 0.25	97.70 ± 0.13	97.56 ± 0.22	97.56 ± 0.22	97.56 ± 0.22
D. melanogaster	97.88 ± 0.20	98.27 ± 0.19	98.07 ± 0.11	98.27 ± 0.11	98.31 ± 0.11	98.29 ± 0.11
A. thaliana	99.05 ± 0.15	99.58 ± 0.05	99.31 ± 0.06	99.51 ± 0.05	99.51 ± 0.05	99.51 ± 0.05
C. elegans	95.12 ± 0.32	98.22 ± 0.24	96.65 ± 0.13	96.92 ± 0.16	96.95 ± 0.16	96.93 ± 0.16

For the *R. norvegicus* organism, the metrics obtained with the TSVM classifier were slightly greater than the results obtained with the ISVM (see Table 1). These results reinforce the hypothesis that the transductive learning model behaves better in scenarios with fewer labelled examples in the training dataset.

In general, the results achieved with ISVM and TSVM classifiers were very similar. That is due to the sequences used in training being well represented for each class. This allows for the inductive model to achieve good results, even with only 10% of the data is used for training.

Fig. 1 (for the *R. norvegicus organism*) shows the distribution of sequences related to the *R. norvegicus* organism, and those is similar to the other organisms investigated (results not shown). It is possible to observe that the proposed methodology generates sequences that are easily separable between the two classes (TIS and nTIS). This suggests that the non-use of the *upstream* conservative positions, and the use of CDSOP nucleotides as nTIS provided an improvement in the performance of the model.

Figure 1: Class separation with proposed methodology



To further explore the performance and the capacity of the TSVM to deal with smaller training sets, two changes were made in Scenario 2. In the first change, only 50% of the available training base was applied. In other words, meaning the 10% of the dataset became 5%, keeping the same data for validation. In the second change, the dataset for training was reduced to 25% of its original size, meaning the 10% of the dataset became 2.5%, keeping the same set for validation.

Table 2: Validation results for scenario 2 - with 2.5% of data

Scenario 2						
Organisms	**Inductive**			**Transductive**		
	Precision	Sensitivity	*F-measure*	Precision	Sensitivity	*F-measure*
R. norvegicus	69.07 ± 25.01	61.90 ± 24.66	63.54 ± 22.21	75.08 ± 4.80	74.97 ± 4.98	75.03 ± 4.89
M. musculus	90.34 ± 2.72	91.87 ± 5.44	90.96 ± 2.48	95.96 ± 1.02	95.97 ± 1.63	95.96 ± 0.82
H. sapiens	94.76 ± 0.20	95.58 ± 0.33	96.63 ± 0.18	96.52 ± 0.23	95.70 ± 0.22	96.71 ± 0.22
D. melanogaster	96.91 ± 0.46	97.83 ± 0.61	97.37 ± 0.12	97.87 ± 0.11	97.91 ± 0.11	97.89 ± 0.10
A. thaliana	98.45 ± 0.27	99.34 ± 0.21	98.89 ± 0.14	99.63 ± 0.09	99.33 ± 0.07	99.49 ± 0.05
C. elegans	93,58 ± 0.18	98.14 ± 0.46	95.80 ± 0.17	97.08 ± 0.19	97.08 ± 0.18	97.08 ± 0.19

The experiment with 5% of the training set showed that, despite reducing the number of training records, the quantity and representativity of data were still sufficient to maintain similar the behaviour for the ISVM and TSVM models. In the experiments with only 2.5% of the training set, it was observed that the TSVM excelled in relation to the ISVM in organisms that have fewer molecules (see Table 2). For the *R. norvegicus* and *M. musculus* organisms, the TSVM results were superior to the ISVM, being on average 5% higher.

To test the behaviour of the ISVM and TSVM models, new molecules, which did not participate in the training and testing processes, were classified by the constructed models. This validation dataset[12] was extracted from *RefSeq* in the period between April 22 and September 22, 2014, except for the organism *C. elegans*, that had 100 molecules collected from April 17 to August 17, 2017. For the last organism, the first 100 molecules that had at least 1081 nucleotides in the CDS region were considered. The models used were the ones that obtained the best results in Scenario 1.

The test results were compared with the tools: *Transdutis-I*, *Transdutis-T* [8], *TISHunter*, *TisMiner* and *NetStart*. The trained models from our work are called *FindTIS-I* (for inductive SVM) and *FindTIS-T* (for Transductive SVM). The results are shown separately by organism. The *C. elegans* organism was evaluated only with the models obtained in this work, since the other methods consider the *upstream* region in their approaches, and this organism has only the CDS region.

The molecules were classified as Hit and nHit, where Hit corresponds to AUG which is TIS, and nHit corresponds to AUG which is TIS but was classified as nTIS. The results obtained show that the proposed methodology in this work is equivalent to the methodology of Pinto et al. [2017]; since in the tests performed, our method obtained the same performance, except for *Transdutis-T* in the *Homo sapiens* and *D. melanogaster* organisms, which slightly outperformed our model (our model missed the prediction in a maximum of 5 molecules). However, it managed to be superior for the *R. norvegicus* organism. For the organism *C. elegans*, our models achieved high success rates, of 99% for ISVM and 98% for TSVM. The other tools do not evaluate this organism because it depends of the *upstream* region.

Table 3: Test results

Organisms	Transdutis-I		Transdutis-T		TISHunter		TIS Miner		NetStart		*FindTIS-I*		*FindTIS-T*	
	Hit	nHit	Hit	nHit	Hit	nHit	Hit	nHit	Hit	nHit	Hit	nHit	Hit	nHit
R. norvegicus	109	16	122	3	112	13	89	36	109	16	123	2	123	2
M. musculus	22	14	36	0	35	1	34	2	31	5	36	0	36	0
H. sapiens	102	11	107	6	106	7	91	22	84	29	102	11	102	11
D. melanogaster	95	11	105	1	93	13	76	30	78	28	102	4	102	4
A. thaliana	15	0	15	0	14	1	12	3	5	10	15	0	15	0
C. elegans	—	—	—	—	—	—	—	—	—	—	97	3	97	3

5 CONCLUSIONS

We carried out a comparison between the inductive and the transductive approaches, using a new methodology for extracting mRNA sequences, in the context of prediction of TIS.

It was possible to verify that, although the *upstream* region is not used in nucleotide extraction, the use of a window that comprehends several *downstream* region positions is able to replace the *upstream* region's characteristics identified by Kozak [1984], regarding the overall prediction performance.

TSVM was shown useful for organisms that have fewer sequenced molecules. It is observed in both Scenarios 1 and 2 that the TSVM was superior for the organism *R. norvegicus*, which has the lower number of molecules among the organisms studied. Also, the use of CDSOP sequences as nTIS in the training process, made the information contained in the CDS region add more features, and improved the performances of both inductive and transductive models.

Considering the experiments with only 2.5% of the training dataset, it was confirmed that the performance of the TSVM is superior to the ISVM in scenarios where the amount of labelled data is scarce.

A limitation of our approach is the following. As the number of unlabelled sequences grow, transductive learning begins to become more and more expensive computationally, since, with each new unlabelled sequences, it is necessary to update the entire model.

To overcome this problem, we intend to apply the cascade strategy together with the transductive learning model. Because transductive learning behaves very well with few labelled sequences, applying it together with the SVM cascade approach may provide promising results. Therefore, it is expected that the models will maintain the quality of the results and have reduced the training times. Also, for future work, we would like to apply this methodology to other organisms.

REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 3, Article 27 (May 2011), 27 pages.

[2] A. Gammerman, V. Vovk, and V. Vapnik. 1998. Learning by Transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98)*. 148–155.

[3] Artemis G. Hatzigeorgiou. 2002. Translation initiation start prediction in human cDNAs with High Accuracy. *Bioinformatics* 18 (2002), 343–350.

[4] Ron Kohavi. 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'95)*. 1137–1143.

[5] Evgeny Kondratovich, Igor I Baskin, and Alexandre Varnek. 2013. Transductive support vector machines: Promising approach to model small and unbalanced datasets. *Molecular Informatics* 32, 3 (2013), 261–266.

[6] M. Kozak. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res* 12, 2 (25 Jan 1984), 857–872.

[7] So Nakagawa, Yoshihito Niimura, Takashi Gojobori, Hiroshi Tanaka, and Kin-ichiro Miura. 2008. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic acids research* 36, 3 (2008), 861–871.

[8] C. L. Pinto, C. N. Nobre, and L. E. Zárate. 2017. Transductive learning as an alternative to translation initiation site identification. *BMC Bioinformatics* 18, 1 (2017), 81.

[9] L. M. Silva, F. C. S. de Teixeira, J. M. Ortega, L. E. Zárate, and C. N. Nobre. 2011. Improvement in the prediction of the translation initiation site through balancing methods, inclusion of acquired knowledge and addition of features to sequences of mRNA. *BMC Genomics* 12, Suppl 4 (22 Dec 2011), S9–S9.

[10] A. Stanescu and D. Caragea. 2015. Predicting cassette exons using transductive learning approaches. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Citseer, IEEE, 1–8.

[11] George Tzanis, Christos Berberidis, and Ioannis Vlahavas. 2006. A novel data mining approach for the accurate prediction of translation initiation sites. In *International Symposium on Biological and Medical Data Analysis*. Springer Berlin Heidelberg, Citseer, Springer Berlin Heidelberg, 92–103.

[12] Were considered, 123 molecules for ***R. norvegicus***, 36 molecules for ***M. musculus***, 113 molecules for ***H. sapiens***, 106 molecules for ***D. melanogaster*** and 15 molecules for ***A. thaliana***

[0] **ACKNOWLEDGMENTS:** We appreciate the financial support given by Foundation for Research Support of the State of Minas Gerais (FAPEMIG) and the Brazilian National Council for Scientific and Technological Development (CNPq).