ISSN 1870-4069

DNA Sequence Recognition using Image Representation

Luis A. Santamaría C., Sarahí Zuñiga H., Ivo H. Pineda T., María J. Somodevilla, Mario Rossainz L.

Benémerita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla, Mexico

aluloc_kapa@hotmail.com, sarahi.zuhe@gmail.com,
ivopinedatorres@gmail.com, {mariasg,rossainz}@cs.buap.mx

Abstract. In recent years, the field of machine learning has progressed enormously in addressing difficult classification problems. The problem raised in this article is to recognize DNA sequences, recognize the boundaries between exons and introns using a graphic representation of DNA sequences and recent methods of deep learning. The objective of this work is to classify DNA sequences using a convolutional neuronal network (CNN). The set of DNA sequences used for the recognition were 1847 sequences from a database with 4 types of hepatitis C virus (type 1, 2, 3 and 6) taken from the repository available on the ViPR page. The other set of sequences used to recognize limits between exons and introns were sequences from the Molecular database (Splice-junction Gene Sequences) Data Set that has 3190 sequences, available on the ICU page, with three classes of sequences: limit exon-intron, limit intron-exon and none. For the processing of the DNA sequences, a representation method was designed where each nitrogenous base is represented in gray scale to form an image. The generated images were used to train the convolutional neuronal network. The results obtained from the CNN trained with the Hepatitis C virus database suggest that the CNNs are suitable for the classification of the images generated from the DNA sequences. This result led us to perform the experiments for the recognition of exons and introns with the UCI database for the recognition of limits between exons and introns. The results obtained were a training precision of 82%, a validation accuracy of 75% and an evaluation accuracy of 80.8%. It is concluded that it is possible to classify the images of DNA sequences of the databases used.

Keywords. Gene recognition, deep learning, convolutional neuronal networks, coding DNA sequences.

1 Introduction

The automatic learning methods allow to identify characteristics that favor the classification, analysis and recognition of patterns. In the area of biology, the use of machine learning methods facilitates the recognition of DNA sequences. This work recognizes the genes of DNA through images. This article is divided into sections, the first section is the state of the art that is the prior knowledge necessary for the recognition of genes. The second section details the methodology that was used for the analysis of DNA sequences. The third and fourth sections show the results and conclusions obtained.

2 State of the Art

The mechanisms or processes of gene prediction are those that, within the area of computational biology, are used for the algorithmic identification of pieces of sequences, usually genomic DNA [8], and that are biologically functional. This, especially includes genes encoding proteins and regulatory sequences. The identification of genes is one of the first and most important steps to understand the genome of a species once it has been sequenced [11].

Deoxyribonucleic acid (DNA) is composed of four molecules called nucleotides or nitrogenous bases: adenine, thymine, guanine and cytosine [9]. A complete DNA molecule or, in other words, a DNA sequence is composed of an alphabet that contains the letters of the four nitrogenous bases.

$$\begin{aligned} \Sigma\{ATGC\}, \\ \phi_i = (V_1, V_2, V_3, \ldots, V_n), \\ V_i \in \Sigma, \end{aligned} \tag{1}$$

where a string ϕ is a sequence of DNA formed elements of the alphabet Σ (equation 1) can define the characteristics of a living organism, containing all the genetic information in inheritance units called genes. The mechanisms or processes of gene prediction are those that, within the area of computational biology, are used for the algorithmic identification of pieces of sequences, usually genomic DNA [1], and which are biologically functional. This, especially includes genes encoding proteins and regulatory sequences. The identification of genes is one of the first and most important steps to understand the genome of a species once it has been sequenced [2].

Splicing junctions are points in a DNA sequence in which “useless” DNA is removed during the process of creating proteins in higher organisms. The problem posed in this data set is to recognize, given a DNA sequence, the boundaries between the exons (the parts of the DNA sequence retained after splicing) and the introns (the parts of the DNA sequence that are cut).This problem consists of two subtasks: recognition of exon/ intron limits (called EI sites) and recognition of intron exon boundaries (IE sites). In the biological community, the limits of IE refer to the “acceptors” while the boundaries of EI are known as “donors” [7]. Both tasks are complicated since there is no standard sequence to recognize introns and exons, for this reason it is interesting to design tools that help us identify and classify them.

The number of research projects on currently valid genomes is increasing at an accelerated rate, and providing a catalog of genes for these new genomes is a challenge. Obtaining a set of well-characterized genes is a basic requirement in the initial steps of any process of creating a genome. Computational gene search methods can be categorized as based on alignment and sequence composition or a combination of both. Methods based on sequence alignment can be used when trying to predict a gene that encodes a protein for which there is a closely related homologue, this is the approach in GeneWise [5] and PROCRUSTES [4].

Algorithms based on sequence composition (also known as gene search methods) contain a probabilistic model of gene structure based on biological signals (splice sites and translation start / stop sites) and composition properties of functional sequences (exons as coding sequences and introns as intermediate sequences between exons and introns). Unlike alignment-based methods, these algorithms rely on the intrinsic properties of genes to construct predicted genetic structures. Genscan [10] and Geneid [3] are the two examples of this approach and can find known genes and new genes as long as the genes conform to the underlying probabilistic model.

As described a DNA strand is a molecule characterized by four nitrogenous bases represented with numerical or alphabetic values: A (adenine), T (thymine), G (guanine) and C (cytosine) [10]. However, the representation of large amounts of information as DNA sequences do not make their mathematical analysis easy, this creates the need to find new ways of representing information.

In Lapedes [6] and his team trained a neural network to recognize genes in DNA sequences, achieved an accuracy of 91.2% in intron / exon splicing junctions and 92.8% in splicing junctions (exon / intron). What gives a motivation to use of convolutional neural networks to solve this same classification problem. This work consisted of looking for a new way to represent DNA sequences for analysis, as has already been mentioned, there are currently different methods to recognize genes, but these representations complicate their analysis. The proposal we present is to generate images from DNA sequences and subject them to analysis with deep learning techniques, specifically convolutional neuronal networks; used for the classification of images. At the moment a mathematical model has not been found that solves the process of classification by neural networks, but its results get to be so high that they surpass 99% in some cases [1].

2.1 Convolutional Neural Network (CNN)

In recent years, the field of machine learning has made tremendous progress in addressing problems of classification, identification and pattern recognition. In particular, it has been found that a type of model called Convolutional Neural Network CNN, which achieves a reasonable performance in tasks of visual recognition, equaling or exceeding human performance in some domains [11]. A CNN is an algorithm for machine learning in which a model learns to perform classification tasks directly from images, videos or sounds. CNNs are especially useful for locating patterns in images in order to recognize objects, faces and scenes. They learn directly from the image data, using patterns to classify images and eliminate the need for a manual features extraction.

Inception-v3 is one of the tools designed for the Visual Recognition challenge. This is a standard task in artificial vision, where the models try to classify complete images in 1000 ImageNet classes. On the other hand, TensorFlow is a tool for automatic learning. Although, it contains a wide range of functionalities, TensorFlow is designed mainly for models of deep neural networks. Modern models of image recognition have millions of parameters; training them from scratch requires a large amount of tagged training data and a large amount of computing power (hundreds of hours of GPU or more). Transfer learning is a technique that cuts much of this by taking a piece from a model that has already been trained in a related task and reusing it in a new model, in Figure 1 an example of a CNN is shown, the filters are applied to each training image with different resolutions, and the output of each convolved image is used as input for the next layer [2]. Although not as accurate in comparison to the full model training, it is surprisingly effective for many applications, works with moderate amounts of training data (thousands, not millions of tagged images) and can be executed in just thirty minutes in one laptop without a GPU [11].

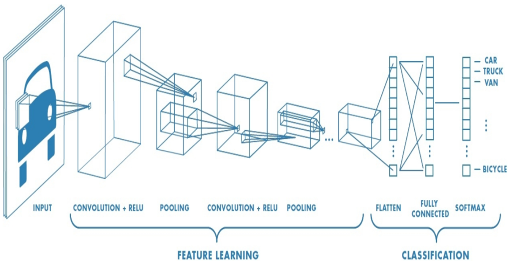


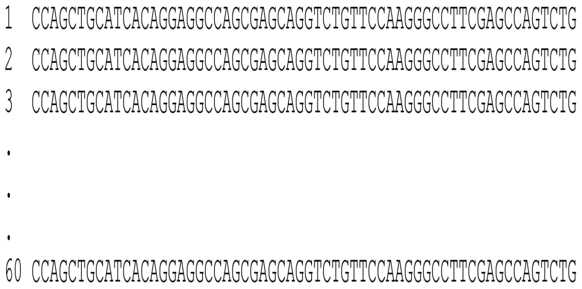
Fig. 1. Example of a Convolutional Neural Network.

3 Methodology

This section describes in detail how images are generated from sequences of DNA and its use to training a convolutional neural network for classification of three classes of sequences. An important appearance that has considered this work is that the CNN are used for the recognition of patterns and classification of images. The sequences of DNA are represented by letters: To use A for the adenine, G for the guanine, C for the cytosine and T for the thymine, however, a CNN is not established to process information under this format, for this reason it has designed a graphic representation of the sequences. The first step was to assign a value of grey to each one of the letters as it shows in Table 1. The scales of grey range from 0 that represents black, to 1 that represents white, in such a way that the resultant intermediate values are tonalities of grey to show a better contrast among them . The second was to do that the sequences could be represented by a specific image to each one. To attain this used a matrix of dimension N X N, where N is the value that coincides with the number of nitrogenous bases (length) of all the sequences of the database employed. Each sequence was planted in the first row and copied in the rest of the rows until having N in total, like this the final result is an image with bars in the scale of grey like which shows in Figure 3, each one of the images obtained is specific for each instance of the database as it observes in figure 2. In total obtained 1847 images of the database of ViPR and 3190 images of the database of the UCI.

Table 1. Color Representation of nitrogenous bases.

Nitrogenous bases	Grey Value
A	0
C	0.3
G	0.7
T	1



```
1  CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
2  CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
3  CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
.
.
.
60 CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
```

Fig. 2. DNA sequencce for coding.

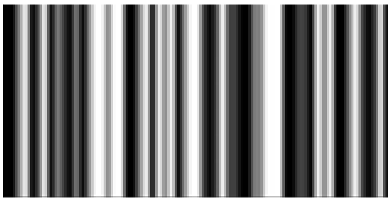


Fig. 3. Image related.

3.1 Use of CNN to DNA Sequences

This subsection describes how a CNN was trained with the representative images of each sequence. An InceptionV3 CNN was used to which the deep learning transfer was applied to categorize the recognition of three classes of DNA sequences: recognition of exon / intron limits (called EI sites), recognition of intron / exon limits (IE sites) and recognition of neither of the previous two (N).

Once it was possible to represent the DNA sequences as images, a CNN was used and with the software libraryTensorFlow was built a classification model based on a pre-trained convolutional neuronal network. Inception V3 CNNs were used to which the deep learning transfer is applied to categorize the recognition of a database with four classes of DNA sequences:Hepatitis C virus type 1, 2, 3 and 6 and the recognition of another database with three classes of exon / intron limits (called EI sites), recognition of intron / exon boundaries (IE sites) and recognition of none of the previous two (N). To adjust the model to our problem, the last layers of the networks are trained with instances obtained from the databases, both networks were trained in 4000 steps.

First the CNN was trained to classify the 4 types of Hepatitis virus, then a CNN was trained with only 2 classes: EI and IE and finally another CNN was trained with all the classes of the database: EI, IE and N to compare the results of the last two neurons.

4 Results

The classification results for the CNN trained with the database of the four types of Hepatitis C virus show a 95 % evaluation accuracy with 145 images tested and at the end of step (k) 4000 the training precision was 94.5 % and validation accuracy of 95 % as seen in the figure 4. The decreasing behavior of entropy during training, is seen in the figure 5 .

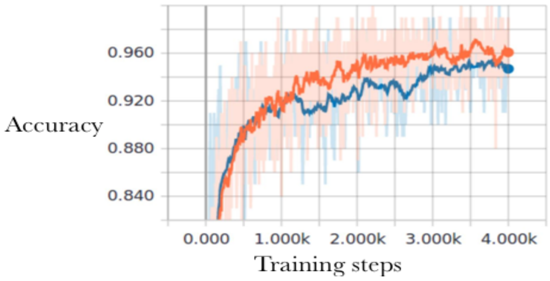


Fig. 4. CNN with classes of Hepatitis C virus type 1, 2, 3 and 6. Orange: training precision. Blue: validation accuracy after 4000 steps (k).

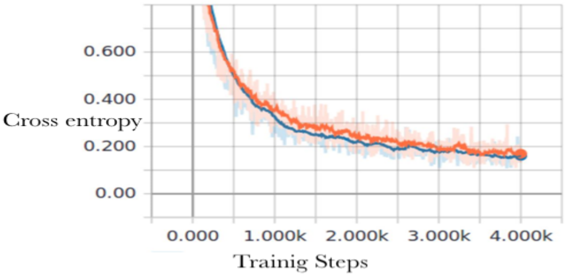


Fig. 5. CNN crossed entropy with classes of Hepatitis C virus type 1, 2, 3 and 6 after 4000 steps (k). Upper stroke: training. Bottom line: validation.

When using a CNN with the classes EI and IE, an evaluation accuracy of 80.8 % is obtained with 177 test images and at the end of step (k) 4000 the training precision is 82 % and the validation accuracy of 75 %. Figure 6 shows how the accuracy of training (orange) and validation (blue) is changing in each step and in Figure 7 shows how entropy decreases with the increase in steps during the training.

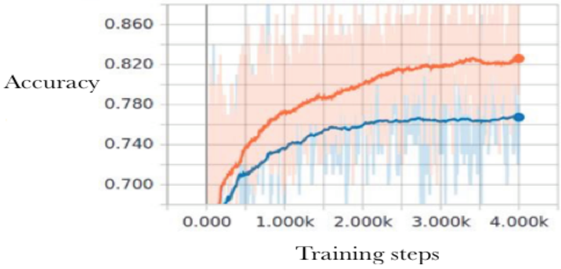


Fig. 6. CNN with classes IE and EI. Orange: training precision. Blue: validation accuracy after 4000 steps (k).

On the other hand, the results of the second CNN where the three classes of the database were used show an evaluation accuracy of 57.5 % with 301 images and at the end of step 4000 the training precision 69 % and the precision of validation with 56 % as shown in Figure 8. Figure 9 shows the changes in entropy at each stage of the training.

sectionConclusions The results obtained from the CNN trained with the Hepatitis C virus database suggest that the automatic learning methodology used in this work is suitable for the classification of the images generated from the DNA sequences, showing important and high percentages of evaluation accuracy, training precision and validation accuracy. These results led us to

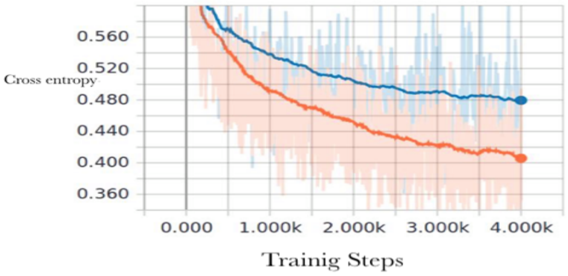


Fig. 7. Cross entropy of CNN with classes IE and EI after 4000 steps (k). Orange: training. Blue: validation.

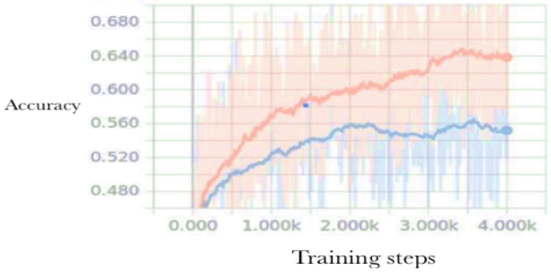


Fig. 8. CNN with classes IE, EI and N. Orange: training precision. Blue: validation accuracy after 4000 steps (k).

carry out the following experiments for the recognition of exons and introns in the following database: For this case, the CNNs show that the percentages of validation accuracy are lower compared to those of a neural network based on the work of Lapedes [6]. The importance of the work is that favorable results are presented to continue exploring the use of convolutional neural networks using the representation of DNA sequences as images, a simple and practical coding method.

In this work, it has been possible to perform DNA sequence classification using a CNN and the results show that CNN are able to perform this classification with an 80.8 % accuracy of evaluation for the experiment with classes IE and EI and 57.5 % for the experiment with classes IE, EI and N. Similar results can be seen in the precision of training and validation of the Figures 6 and 8. In the case of the four types of hepatitis, results of up to 94.5 % of evaluation accuracy are achieved.

The difference between the results obtained for the experiments with two and three classes can be justified that increasing the number of classes increases the entropy Figures 7 and 9. Cross entropy is a metric that can be used to reflect the accuracy of probabilistic forecasts and is closely linked to the maximum likelihood estimation. The crossed entropy is a function that allows to evaluate

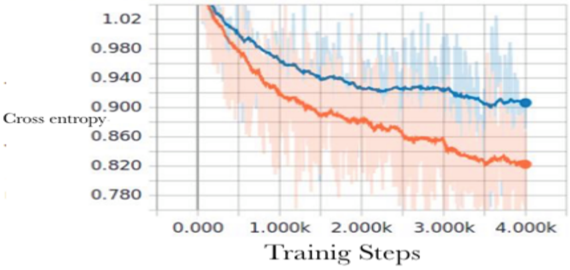


Fig. 9. Cross entropy of CNN with classes IE, EI and N after 4000 steps (k). Orange: training. Blue: validation.

the result of the classification instead of using the metric of the mean square error, the value of the crossed entropy allows to evaluate the progress of the process of learning of the information [1].

On the other hand, there is talk that the transfer of learning is good when there are few images to train the network and that allows to reach acceptable results in most cases, however, it is still possible to further improve validation and training accuracy and decrease entropy if a neural network is trained from scratch,that is, we must have a database of millions of instances and a computer with GPU to train this network but it will surely offer better results than the pre-trained CNN we use for this work.

In conclusion, it can be stated that a convolutional neuronal network of the InceptionV3 model is capable of classifying DNA sequences if the sequence is processed and transformed into an image, however, the accuracy percentages can be improved if a CNN is trained with a larger sequence base.

Acknowledgments

The authors thank the National Council of Science and Technology (CONACyT) of México, the Benemérita Universidad Autónoma de Puebla, which through the Faculty of Computer Science have provided the necessary support for the realization and presentation of this work.

References

1. How to retrain an image classifier for new categories. `https://www.tensorflow.org/tutorials/image_retraining`, accessed: 2018-05-28
2. Mathworks (2018). deep learning. `https://la.mathworks.com/solutions/deep-learning/convolutional-neural-network.html`, accessed: 2018-05-27
3. Almeida, J.S., Vinga, S.: Universal sequence map (usm) of arbitrary discrete sequences. BMC bioinformatics 3(1), 6 (2002)
4. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic dna1. Journal of molecular biology 268(1), 78–94 (1997)

5. Gelfand, M.S., Mironov, A.A., Pevzner, P.A.: Gene recognition via spliced sequence alignment. Proceedings of the National Academy of Sciences 93(17), 9061–9066 (1996)
6. Lapedes, A., Barnes, C., Burks, C., Farber, R., Sirotkin, K.: Application of neural networks and other machine learning algorithms to dna sequence analysis. Tech. rep., Los Alamos National Lab., NM (USA) (1988)
7. Noordewier, M.O., Towell, G.G., Shavlik, J.W.: Training knowledge-based neural networks to recognize genes in dna sequences. In: Advances in neural information processing systems. pp. 530–536 (1991)
8. Ouzounis, C.A.: Rise and demise of bioinformatics? promise and progress. PLoS computational biology 8(4), e1002487 (2012)
9. Panduro, A.: Biología molecular en la clínica. McGraw-Hill Interamericana (2009)
10. Parra, G., Blanco, E., Guigó, R.: Geneid in drosophila. Genome research 10(4), 511–515 (2000)
11. Salzberg, S., Searls, D., Kasif, S.: Computational gene prediction using neural networks and similarity search. Computational Methods in Molecular Biology 32, 109 (1998)