BITS :: Call for Abstracts 2019 - Oral communication

Type	Oral communication
Session	Machine Learning in Bioinformatics
Title	A Deep Neural network for extracting features from DNA nucleosome-forming sequences
All Authors	Domenico Amato(1), Giosuè Lo Bosco(1), Riccardo Rizzo(2).

Affiliation

(1)Dipartimento di Matematica e Informatica, Università degli studi di Palermo, Via Archirafi, 34, 90123 Palermo, Italy

(2)CNR-ICAR, Consiglio Nazionale delle Ricerche, Via Ugo La Malfa, 153, 90146 Palermo, Italy.

Motivation

Nucleosomes are the fundamental sub-unit of chromatin. They are composed by a histone octamer wrapping ~147 pairs of DNA bases. There is evidence of sequence specificity and periodicity of nucleosomes in specific organisms [1]. Such kind of studies has open the possibility to use standard machine learning algorithm in order to automatically predict the presence of nucleosomes by solely using the DNA sequence. In general, being DNA sequences text strings, any machine learning method processing such kind of input must be submitted to a feature representation phase with the main purpose of transforming the string into a numerical vector. The choice of the representation is crucial for the success of the Machine Learning method, for this reason, problem knowledge by experts is usually adopted. In this work, we avoid the feature representation step by using the so-called Deep Neural Networks. They are based on specific architectures able to correctly identify the most suitable space for representing the data. Recently we have introduced a deep learning model based on a Convolutive layer (CL) followed by a Recurrent layer (DLNN) for nucleosome classification [2]. We adopt it in this work as a feature extractor. This is realized after the learning phase of the network, using the CL as input for Support vector machine (SVM), logistic regression (LR) and random forest (RF) classifiers. The area under the Roc curves (AUC) of nucleosome vs no nucleosome sequences taken from Yeast, Drosophila and Human species, are carried out for each of the used classifiers and compared with the DLNN classifier.

Methods

The DLNN we propose is composed by a Convolutional layer (CL), a special kind of recurrent layer called long short term memory layer (LSTM) and two densely connected layers (DL) (see Figure 1 for details). We have used sequences related to nucleosome presence and absence coming from different species, Yeast (YS), Drosophila (DM) and Human (HM). Sequences have been transformed following one hot-encoding scheme, where the categories are the DNA nucleotides. A single input is thus represented as a matrix of 4xL binary values (L=147bp). The CL uses a kernel of length 3 and is formed by 50 neural units, its role is to provide a feature extractor for each input data. Its output is used by an LSTM layer and is needed to learn periodic patterns of features that characterize the nucleosome class. Finally, the DL layers perform the classification. To measure the efficiency of CL feature extractor alone, we use its output as input for SVM, LR, RF classifiers, after the learning phase of the whole DLNN. In particular, the extracted features corresponds to the concatenation of the outputs of the 50 neural units of CL.

Results

Results have been computed using 10-fold cross-validation. For each fold, a ROC curve has been calculated and the AUC (Area under the curve) value has been used as evaluation measure. Table 1 shows an overview of the results. The DLNN network used as a classifier (CL+LSTM+2DL) whose AUC values are in the column named LSTM, obtains the best results with an average AUC of 90% for Yeast and Human and 71% for Drosophila (standard deviation below 1%). The other classifiers, when using the CL layer of DLNN as feature extractor, seems to have not enough efficacy. Note that the main difference between DLNN and the other classifier is the presence of an LSTM layer that provides the specific property of taking into account the nucleotide order in the nucleosome sequence. This is a property that has been found on the YS organism [1]. Our study seems to confirm a periodicity also in other species such as DM and HM. However, the results of SVM, LR, RF methods are in accordance with those obtained by other methods present in the literature, which uses instead a feature engineering process [3]. This demonstrates that the convolutional layer of a DLNN network is capable of extracting useful features for the case of nucleosome classification.

Info

-

Figure

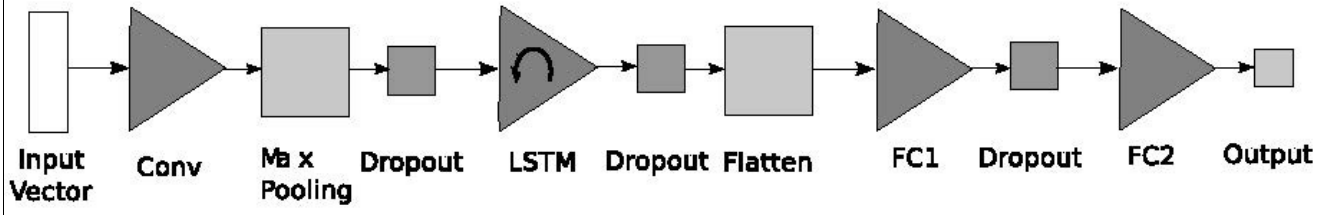

Figure 1: The network architecture

	LSTM	SVM - RBF	SVM - Linear	Logistic Regression	Random Forest
Yeast	0.92	0.87	0.56	0.77	0.76
Drosophila	0.71	0.65	0.56	0.71	0.67
Human	0.9	0.84	0.51	0.67	0.81

Table 1: AUC values of the used classifiers

Availability	http://www.math.unipa.it/lobosco/private/Abstract_BITS_Amato_Lo_Bosco_Rizzo.pdf

Corresponding Author

Name, Surname	Giosue', Lo Bosco
Email	giosue.lobosco@unipa.it

Submitted on	26.04.2019