Predicting the oncogenic potential of gene fusions using Convolutional Neural Networks

Marta Lovino[1 [0000−0001−7124−8319]], Gianvito Urgese[1 [0000−0003−2672−7593]], Enrico Macii[2 [0000−0001−9046−5618]], Santa di Cataldo[1 [0000−0002−6239−8945]], and Elisa Ficarra[1 [0000−0002−8061−2124]]

[1] Politecnico di Torino, Dept. of Control and Computer Engineering, Corso Duca Degli Abruzzi 24, 10129, Torino,Italy {marta.lovino, gianvito.urgese, santa,dicataldo, elisa.ficarra}@polito.it

[2] Politecnico di Torino, Interuniversity Dept. of Regional and Urban Studies and Planning, Corso Duca Degli Abruzzi 24, 10129, Torino, Italy enrico.macii@polito.it

Abstract. Predicting the oncogenic potential of a gene fusion transcript is an important and challenging task in the study of cancer development. To this date, the available approaches mostly rely on protein domain analysis to provide a probability score explaining the oncogenic potential of a gene fusion. In this paper, a Convolutional Neural Network model is proposed to discriminate gene fusions into oncogenic or non-oncogenic, exploiting only the protein sequence without protein domain information. Our proposed model obtained accuracy value close to 90% on a dataset of fused sequences.

Keywords: Gene Fusions · Deep Learning · Convolutional Neural Networks.

1 Scientific Background

Nowadays, the increased availability of Next Generation Sequencing (NGS) data enables new unforeseen insights into the relation between some genetic rearrangements and cancer development. In this regard, a challenging area is represented by the study of gene fusions, a genetic aberration where two separate DNA regions (usually two distinct genes) join together into a hybrid gene. The genes retained at 5p' and 3p' of the fused sequence are conventionally called 5p' gene and 3p' gene, respectively. If the promoter region of at least one of the two genes is retained in the fusion, the erroneous sequence is transcribed at the RNA level, and the aberrated transcript can result into an abnormal protein [7].

Since the discovery of the first genetic rearrangement by Nowell and Hungerford in 1960, a large number of gene fusions have been associated to cancer development and used as cancer predictors [7]. However, gene fusions do not automatically relate to carcinogenic processes, as they can be found in large number even in non-tumoral samples [2]. In light of the above, predicting whether an

aberrated transcript will result into a functional protein or a cancer driver is a very critical and challenging task in the study of cancer development.

To the best of our knowledge, all current approaches reconstruct the candidate fusion from original sequenced data and apply different types of machine learning methods to perform protein domain analysis. For example, the tools Oncofuse [10] and Pegasus [1] use respectively naive Bayes Network and decision tree classifiers to provide an oncogenic probability score for the fusion based on protein domain data.

To this date, a large number of machine learning approaches have been proposed to solve different types of DNA sequence classification problems, with an increasing trend in the use of deep learning techniques [8, 9]. More specifically, Convolutional Neural Networks (CNNs), a class of deep, feed-forward neural networks originally designed for image classification problems, are now exploited in many DNA sequence analysis tasks for their ability to automatically learn the features from the training data, avoiding the design of handcrafted descriptors. Among the many tasks, CNNs have been successfully applied to model the properties and functions of DNA sequences, to the prediction of single-cell DNA methylation states and microRNA targets, as well as to the recognition of splice junction sites and promoter sequence regions [8].

In this work we exploit CNN to classify candidate gene fusions into cancer driving and non-carcinogenic fusions, outputting a categorical class label instead of a probability score. Unlike previous approaches, our model exploits human reference sequences (and not original sequencing data), relying only on the fusion sequence, with no additional input about conserved or lost protein domains. By doing so, our aim is to avoid any possible bias that the prediction models leveraged by protein domain analysis may introduce into the classification task, as well as to improve the generalization capabilities and ease-of-retraining of the classifier. This is a very important trait in a continuously-evolving field of knowledge such as the study of cancer development.

To design a completely protein domain independent model, we provide the real amino acid composition of the fused protein to the network, without any other additional data interpretation.

2 Materials and Methods

As already mentioned, the purpose of our work is to discriminate between gene fusions with functional oncogenic potential (referred to as *Onco class*) and fusions that are not involved in a carcinogenic process (referred to as *NotOnco class*), without any previous information on the protein domains retained or lost in the fusion sequences. For this purpose, we exploit the ability of the CNN to recognize local spatial patterns that are significant for the classification without requiring any a priori feature description of the two classes.

Overall our dataset contains a total number 1741 reconstructed fused sequences, respectively 1005 for the *Onco class* and 736 for the *NotOnco class*. As CNNs traditionally take images as input, we apply and compare three different

encoding methods to transform fusion sequences into image-like data structures. The process of data retrieval, encoding from sequence to images and CNN design and training are described in the following.

2.1 Fusion data retrieval

Gene fusion data were retrieved from two different sources, respectively for the *Onco* and the *NotOnco class*.

Cosmic, a catalogue for somatic mutations in cancer [6], was used for the *Onco class*. This catalogue provides per each fusion the transcript name of both 5p' and 3p' genes, as well as breakpoint information on the retained transcripts considering UTR regions. For our work we selected only the coding sequence (CDS) retained in the fusion. As this sequence translates into a protein which may or may not be involved in an oncogenic process, it is the only information that is significant for our classification task. For consistency with the *NotOnco class* data, we reconstructed a total number of 1011 fusion sequences from the GRCh37 version of the catalogue.

Data for the *NotOnco class* were reconstructed based on Babicenau at al. work on recurrent chimeric fusion RNAs in non-cancer tissues and cells [2]. In this work, SOAPfuse (a tool for gene fusion analysis) was applied on 171 non-neoplastic tissue samples from 27 different tissues, identifying 291 recurrent fusions (i.e. fusions that are detected in more than one sample) involving 238 gene pairs. Per each of these fusions authors report the breakpoint position on human reference genome hg19 of both fused genes. As no information is provided about which part of the transcript is retained in the fusion, we assumed the most common configuration, where the fused sequence is the result of the region near the promoter for the 5p' gene, and of the ending region for the 3p' gene. This assumption is biologically consistent, since a fused transcript needs a promoter region to be translated into protein, and has no impact to our classification task. As a matter of fact, we observed that 91% of the *Onco class* fused transcripts included the region near the promoter for 5p' gene and the ending region for the 3p' gene. On top of that, when selecting the proper CDS region according to the above mentioned configuration, the same CDS region may be involved in more than one trascript. Therefore, we decided to consider as *NotOnco class* all the fusion sequences resulting from all possible combinations of transcripts at 5p' gene with transcripts at 3p' gene. In order to avoid any biases, we discarded all the cases where the intron can be retained in the fusion trascripts. This led to obtain for the *NotOnco class* a total number of 741 fusions which involve 524 transcripts.

Three transcripts are present in both the *Onco class* and the *NotOnco class*.

2.2 Encoding: from sequences to images

Once all the fused sequences had been reconstructed, they were translated into protein sequences following the Amino Acid Translational Table. The translation process is in-frame, because transcripts were taken from the beginning of the

coding sequence identifiable by the ATG triplet, (a.k.a. initiation codon). As CNN are inherently designed to take images as input, the fused amino acid sequence needs to be converted into a $N \times M \times C$ data structure, where N and M are length and width of the image and C the number of channels. For our purposes, N was set to 3000. Hence, we discarded longer sequences and padded the shorter ones using a fake amino acid. By doing so, we obtained a total number of 1741 strings (1005 for the *Onco class* and 736 for the *NotOnco class*, respectively) of 22 different letters, each corresponding to one amino acid (21 real amino acids plus the fake one).

Popular methods for string encoding are ordinal encoding and one-hot encoding, eventually with some variations.

Ordinal encoding substitutes the i^{th} letter in a fusion with a fixed value corresponding to a unique amino acid. Hence, the resulting image will have minimal dimensions $N = 3000 \times M = 1 \times C = 1$, with memory saving advantages compared to other techniques. On the other hand, the incremental values assigned to the amino acids establish an artificial ordering which may bias the representation [5].

One-hot encoding assigns to the i_{th} letter a vector of length L, where each j_{th} element corresponds to a feature. In standard one-hot encoding features are the amino acids: hence, the i_{th} letter is encoded by a vector of all zeros, except for the j^{th} element associated to the amino acid, which is set to 1.

In our work we explored yet another encoding solution, with features corresponding to 28 real amino acid properties (i.e. hydrophobicity, ionic, mass, polarity, etc.) taken from Bulka's work [3]. Hence, in the following we will refer to this strategy as *Bulka's encoding*. In case of on/off properties, the j_{th} element is set to 0 or to 1, based on the fact that the amino acid has or does not have that specific property. For the other ones that are not on/off (i.e. number of H bonds, isoelectric point and hydrophobicity) it is set to the normalized value of that property. For both one-hot and Bulka's encoding strategies, the size of the obtained images is $N = 3000 \times M = 1 \times C = L$.

As the CNN model will inherently assume spatial correlations between adjacent pixels, the data structure was arranged so that the amino acid features constitute the third dimension (i.e. channels) of the image.

Overall the encoding step is summarized in the first section of Fig. 1.

2.3 CNN architecture and training paradigm

As shown in the second section of Fig. 1, we designed a CNN model with two convolutional layers (kernel size 5) followed by two max pooling layers (kernel size 2). To avoid overfitting, we set dropout to 0.5 and learning rate to 0.01. After flattening, we inserted a 1000-units dense layer with ReLU activation function and a final single unit dense layer with sigmoid activation function, which provides the classification output. Batch size was set to 256 and number of epochs to 50, and the network was trained by backpropagation implementing a Stochastic gradient descent optimizer.

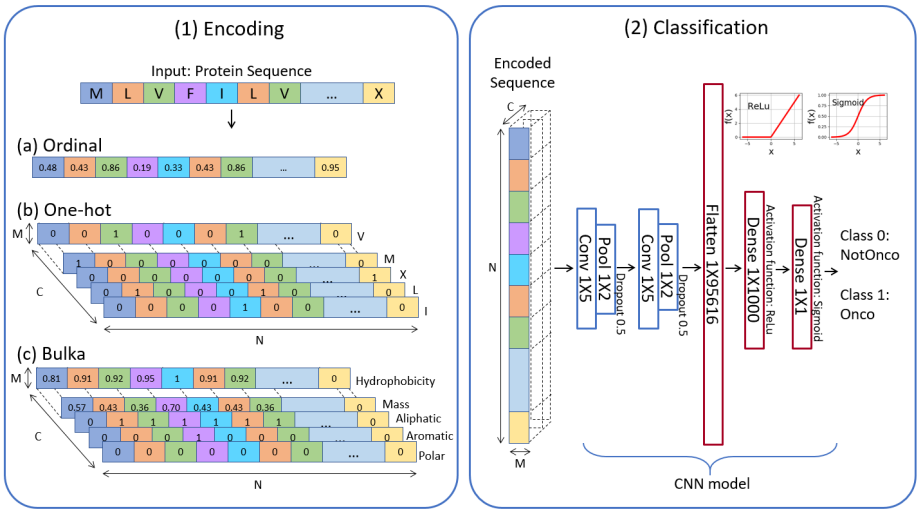


Fig. 1. Overview of the encoding and classification process.

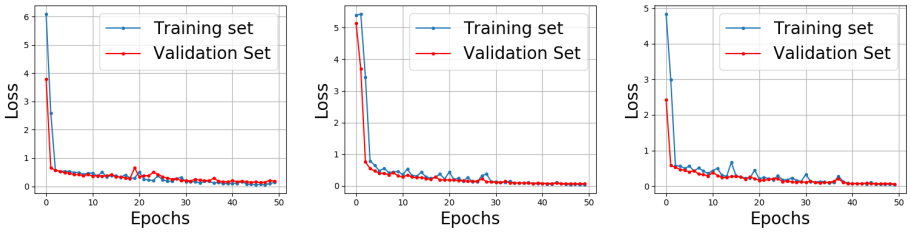
The CNN was implemented in Keras python library under Tensorflow backend [4].

3 Results

A first set of experiments aimed at evaluating the performance of the network in the classification of completely new fused transcripts, using different types of encoding techniques. For this purpose, we created a random partition of the available dataset, where we ensured a complete independence of the training and test sets in terms of involved transcripts. More specifically, we included in the test set only fused sequences whose 5p' and 3p' genes were both not present in any of the fused sequences used for training. This configuration resulted in 1490 samples for the training set and 251 samples for the test set, respectively. With these sets, we trained and tested our CNN model with the three different encoding methods (i.e. ordinal, standard one-hot and Bulka's encoding). In order to asses the stability of the network in terms of independence from weights initialization, we trained and tested the model five times per each type of encoding. As it is visible from the trend of the loss functions during the test set, shown in Fig. 2, the network converged well within 50 epochs.

The test accuracy values obtained in the five runs per each type of encoding are shown in the form of box-plots in Fig. 3, with black boxes ranging from the 25% to the 75% percentile of the accuracy values, and red lines indicating the median accuracy value over the five runs.

From the plots in Fig. 3 we can make the following observations. i) Ordinal encoding consistently achieved the lowest accuracy (around 52%, with very low



(a) (b) (c)

Fig. 2. Loss functions of CNN models using different encodings: 2(a) ordinal encoding, 2(b) standard one-hot encoding and 2(c) Bulka's encoding.

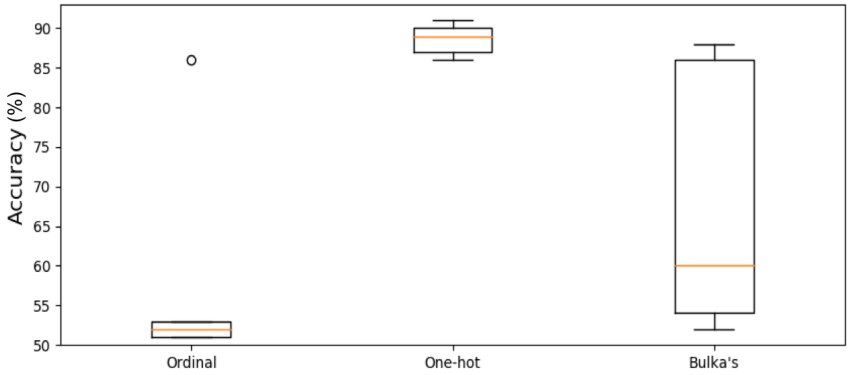


Fig. 3. Box-plot of test accuracy values over five runs, using three different types of encoding techniques.

variation over the five runs). ii) Bulka's encoding obtained on average higher accuracy than ordinal encoding (median accuracy value around 60%), but at the price of a very high variability of the results (Bulka's box ranges from 52% to 88% accuracy). iii) Standard one-hot encoding had a very good accuracy (median value 89%), coupled with reasonably low variability.

Based on our results, standard one-hot encoding provided the best compromise, in terms of classification accuracy and stability of the model. This evidence can be explained by taking into consideration the three different encoding designs. On one hand, ordinal encoding introduces a very strong bias into the representation, because it forcefully creates an alphabetical ordering of the 22 amino acids. This easily explains the low accuracy values obtained by this type of encoding. On the other hand, Bulka's encoding uses physical properties of the proteins to univocally represent each amino acid, without implying any type of ordering. Nonetheless, there is no certainty about the significance of the specific properties that were chosen for the representation, nor of their complete independence. This might explain the high instability of the classification model

leveraging upon Bulka's technique. In the end, according to our results, standard one-hot encoding ensures the most unbiased data representation, and hence the highest classification accuracy. Based on this evidence, we selected this type of encoding for our next set of experiments.

Because of the limited number of transcripts involved in the dataset, one might argue that the high accuracy of the network derives from a sort of memorization of the transcript sequences, and not from a real capability of discriminating significant patterns on the input data. To prove this hypothesis wrong, we performed a second set of experiments, giving the entire transcripts of both the *NotOnco* and the *Onco class* as input to our CNN model. The rationale of this experiment is to ensure that the network does not blindly assign all the first set of transcripts to the *NotOnco class* and all the second set of transcripts to the *Onco class*, respectively.

As a result of this experiment, we obtained that only 70% of the first set of transcripts were classified as *NotOnco class* and the 78% of the second set as *Onco class*, respectively. As a fair amount of the whole transcripts were still assigned a class that is different from the one they were extracted from, we can reasonably conclude that the classification task is not driven by the transcript sequence alone.

4 Conclusion

In the end, our experiments proved that the proposed CNN approach is able to predict the oncogenicity of a gene fusion with a satisfactory level of accuracy, relying only on the fused sequence with no additional information. Tests on three different encoding methods demonstrated that standard one-hot encoding was the most suitable for the representation of the amino acid sequence.

Future works will focus on the interpretation of the features extracted by the locally connected stages of the CNN, in order to obtain a deeper understanding of the specific biological patterns that mostly influence the carcinogenic potential of a gene fusion. On top of that, we plan to increase as much as possible the number of samples used to train the CNN, with the aim of improving the generalization capabilities of our model.

References

1. Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C.H., Frattini, V., Lasorella, A., Iavarone, A., Inghirami, G., Rabadan, R.: Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. BMC systems biology **8**(1), 97 (2014)
2. Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., et al.: Recurrent chimeric fusion rnas in non-cancer tissues and cells. Nucleic acids research **44**(6), 2859–2872 (2016)
3. Bulka, B., Freeland, S.J., et al.: An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices. BMC bioinformatics **7**(1), 329 (2006)

4. Chollet, F., et al.: Keras. https://keras.io (2015)
5. Choong, A.C.H., Lee, N.K.: Evaluation of convolutionary neural networks modeling of dna sequences using ordinal versus one-hot encoding method. bioRxiv p. 186965 (2017)
6. Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al.: Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic acids research **39**(suppl_1), D945–D950 (2010)
7. Mertens, F., Johansson, B., Fioretos, T., Mitelman, F.: The emerging complexity of gene fusions in cancer. Nature Reviews Cancer **15**(6), 371 (2015)
8. Min, S., Lee, B., Yoon, S.: Deep learning in bioinformatics. Briefings in bioinformatics **18**(5), 851–869 (2017)
9. Rizzo, R., Fiannaca, A., La Rosa, M., Urso, A.: A deep learning approach to dna sequence classification. In: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. pp. 129–140. Springer (2015)
10. Shugay, M., Ortiz de Mendíbil, I., Vizmanos, J.L., Novo, F.J.: Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics **29**(20), 2539–2546 (2013)