Journal of Biomedical Informatics 129 (2022) 104057

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Identifying the oncogenic potential of gene fusions exploiting miRNAs

Marta Lovino [b,*], Marilisa Montemurro [a], Venere S Barrese [a], Elisa Ficarra [b]

[a] Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, Italy
[b] University of Modena and Reggio Emilia, Via Vivarelli 10/1, 41125 Modena, Italy

ARTICLE INFO

Keywords:
Gene fusions
Deep learning
Oncogenic potential
Transcription factor
Gene ontologies
miRNA

ABSTRACT

It is estimated that oncogenic gene fusions cause about 20% of human cancer morbidity. Identifying potentially oncogenic gene fusions may improve affected patients' diagnosis and treatment. Previous approaches to this issue included exploiting specific gene-related information, such as gene function and regulation. Here we propose a model that profits from the previous findings and includes the microRNAs in the oncogenic assessment. We present ChimerDriver, a tool to classify gene fusions as oncogenic or not oncogenic. ChimerDriver is based on a specifically designed neural network and trained on genetic and post-transcriptional information to obtain a reliable classification.

The designed neural network integrates information related to transcription factors, gene ontologies, microRNAs and other detailed information related to the functions of the genes involved in the fusion and the gene fusion structure. As a result, the performances on the test set reached 0.83 f1-score and 96% recall. The comparison with state-of-the-art tools returned comparable or higher results. Moreover, ChimerDriver performed well in a real-world case where 21 out of 24 validated gene fusion samples were detected by the gene fusion detection tool Starfusion.

ChimerDriver integrates transcriptional and post-transcriptional information in an ad-hoc designed neural network to effectively discriminate oncogenic gene fusions from passenger ones. ChimerDriver source code is freely available at https://github.com/martalovino/ChimerDriver.

1. Background

Gene fusions are one of the most common somatic mutations and are considered to be responsible for 20% of global human cancer morbidity [11,36]. A gene fusion is a biological event where two independent genes fuse to form a hybrid gene. In the most common case, one gene retains the promoter region and the other one provides the end of the hybrid gene. The former is called 5p' gene, while the latter is called 3p' gene. The position where the break occurs is called breakpoint.

The advent of next-generation sequencing (NGS), the spread of machine and deep learning in bioinformatics [4,28,30,39] and the development of fusion detection algorithms [10,20,21,32] led to the discovery of hundreds of novel fusion sequences.

However, not all gene fusions are oncogenic. Indeed, some are genuinely expressed in normal human cells [15] or constitute passenger events [44]. At the same time, other gene fusions are considered to be responsible for a significant percentage of specific kinds of tumors [23,26,33,45].

A precise diagnosis of oncogenic gene fusions can inform therapeutics treatments [7,42] and be used to predict prognosis, patient survival, and treatment response [11]. Additionally, focusing the research on a smaller number of putative oncogenic fusions a diagnosis could take less time; thus, the risks related to misdiagnosis and waiting may be significantly reduced for the patients.

However, discriminating between cancer-driver fusions and non-driver events is not a trivial task.

The first necessary step to solve this problem is performed by the fusion detection tools [20,21,32], that identify the candidate gene fusions relying on the sample's reads, trying to reduce as much as possible the number of false positives (i.e., detected gene fusions that are not found in the sample in later lab validation). Additional studies proposed more sophisticated approaches based on machine learning (ML) techniques applied to the output of fusion detection tools. Specifically, Oncofuse [34] and Pegasus [14] are noteworthy and use protein domains of the fusion proteins to train the models and predict the oncogenic potential of a fusion. Undoubtedly protein domains are highly

[*] Corresponding author.
E-mail address: marta.lovino@unimore.it (M. Lovino).

https://doi.org/10.1016/j.jbi.2022.104057
Received 30 November 2021; Received in revised form 14 March 2022; Accepted 15 March 2022
Available online 24 March 2022

informative for the characterization of gene fusions. However, using such information as a feature for the ML model requires careful processing from scratch whenever the training database is updated with novel validated fusions.

Recently, previous works explored deep-learning (DL) techniques [27] and presented DEEPPrior [29], a DL model to perform gene fusion prioritization using amino acid sequences of the fusion proteins, based on a Convolutional Neural Network (CNN) and a bidirectional Long Short Term Memory (LSTM) network. Compared to the state-of-the-art tools, this approach is highly effective in accomplishing the classification task with the advantage of avoiding labor-intensive processing of the protein domains.

However, it is known that the oncogenic potential of a molecule depends not only on the sequence itself but also on the effect of post-transcriptional regulatory processes[12].

Transcription Factors (TFs) and micro-RNAs (miRNAs) play a decisive role in the transcriptional and post-transcriptional regulatory processes [31] and can contribute to determining the gene fusion outcome.

To date, most of the available tools exploit transcriptional information and common gene properties to accomplish this task without considering the post-transcriptional regulators affecting the oncogenic processes.

Here, we present ChimerDriver, a new DL architecture based on a Multi-Layer Perceptron (MLP) that integrates gene-related information with miRNAs and TFs, including then in the model transcriptional and post-transcriptional regulative information. Indeed, ChimerDriver exploits the knowledge about TFs and miRNAs targetting each of the genes involved in the fusion to perform gene fusion classification.

ChimerDriver was tested on multiple publicly available datasets and exhibited better classification performance with respect to the state-of-the-art tools. In the end, post-transcriptional regulators confirm the central role in discovering oncogenic processes and miRNAs; in particular, they are a precious source of information to improve the prediction of the oncogenic potential of gene fusions.

In the following, a detailed description of model, its architecture and the input datasets is provided into the Material and methods section. Results are illustrated in Results section. The discussion and conclusion are reported in Discussion and Conclusions sections, respectively.

2. Material and methods

This section introduces the proposed pipeline for the classification of gene fusions. In detail, after a brief overview of ChimerDriver architecture, it illustrates the classification model, the training and testing sets, and the model input features.

2.1. ChimerDriver architecture

Fig. 1 shows the conceptual schema of ChimerDriver. The tool is made of three main modules: the data integration module, which is in charge of extracting the model input features from the input data and integrating them; the feature selection model, which reduces the number of input features through a Random Forest; the classification module, which performs gene fusion prioritization using a neural network. The adopted classification model, the training and testing sets, and the input features are discussed in the following sections.

2.2. Model design

ChimerDriver classifies gene fusions using a Multi-Layer Perceptron (MLP). MLPs are a classical type of feed-forward neural network that, thanks to their flexibility, may be applied to multiple types of data and learn non-linear correlations among them, even when they are produced from different sources [18,38,40]. Besides, the performances achieved by the MLP are, on average, higher than those achieved by traditional machine learning methods on this task (refer to Supplementary Method 1 and Supplementary Table 2, in Supplementary Materials, for more details). For these reasons, the selected model to evaluate the oncogenic potential of the gene fusions is a Multi-Layer Perceptron.

According to grid search results, the best network configuration was denoted by four layers with 512, 256, 128, 64 nodes, all characterized by the tanh activation function, and with a learning rate and dropout values equal to 0.01 and 0.2, respectively. In order to prevent the model overfitting, the early stopping technique has been applied. Please refer to Supplementary Table 1, in the Supplementary Materials, for additional details about the model architecture.

2.3. Dataset

We carefully designed a dataset to train and test our tool. Specifically, the training set consisted of 1765 gene fusion samples: 1059 were oncogenic, and the remaining 706 were not oncogenic. The oncogenic samples were extracted from COSMIC (Catalog of Somatic Mutations in Cancer), a popular database containing information about gene fusions involved in solid tumors and leukemias [13]. Besides, chosen oncogenic gene fusions were already experimentally validated. Finally, the 706 not-oncogenic gene fusions were reported by Babicenau et al. [3] and detected by a gene fusion detection tool in non-neoplastic tissues.

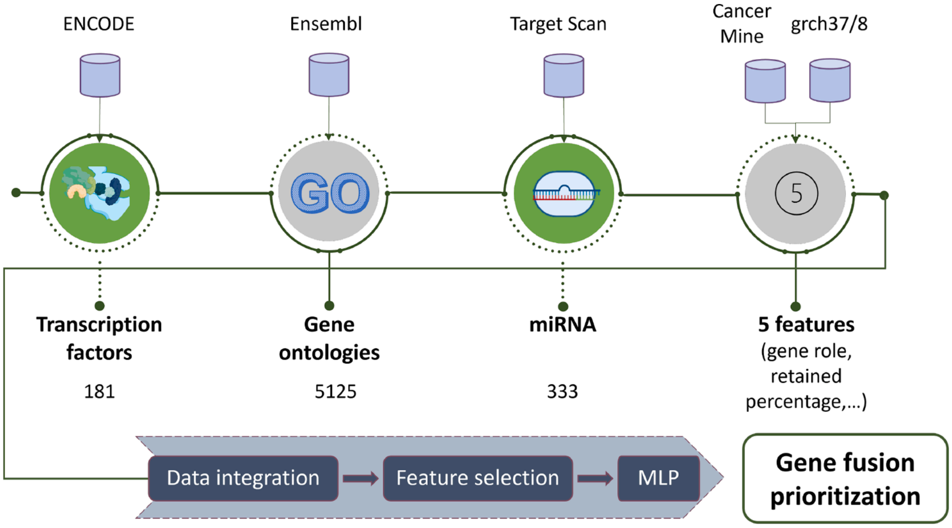


Fig. 1. Conceptual schema of ChimerDriver architecture.

The testing set consisted of 2623 oncogenic gene fusions and 2254 not-oncogenic gene fusions. In detail, for the positive samples, we used the database provided by Gao et al. [17], which results from the application of three fusion detection tools on the TCGA database. Upon request, the authors kindly provided validated gene fusion samples, for which WGS data were available. From this collection, we extracted 2622 oncogenic gene fusions. In addition, we incorporated 2254 not-oncogenic gene fusions found in healthy tissues and reported by Babicenau et al. [3].

Finally, to avoid overfitting, we imposed that the genes involved in the training set gene fusions were not present in the tested gene fusions. This distinction allowed us to verify that the model is sufficiently robust and has learned the oncogenic characteristics of the gene fusions and not specific information relating to the individual genes.

2.4. *Input features*

The model input features were selected from multiple sources to assess different gene fusions' characteristics.

The first five features are obtained from the gene fusion structure, and from Cancermine [22], a literature-mined database of drivers, oncogenes, and tumor suppressors in cancer. In detail, given the breakpoint coordinates, two features correspond to the retained percentage of 5p' and 3p' genes in the gene fusion. One additional feature analyzes the strands of 5p' and 3p' genes, and it is equal to 1 if the two strands are concordant (i.e., the two genes transcribe in the same direction), 0 otherwise. The remaining two features correspond to the nature of each gene according to Cancermine [22]: 'Oncogenic', 'Driver', 'Tumor suppressor' or 'Other' when none of the above options applies. This feature contributes to assessing the functional profiling of the gene fusion.

We added TFs and GOs involving the fused genes to the aforementioned input features. In fact, multiple studies [6,34,47] demonstrated that using these molecules in the gene fusion classification task has a positive impact on the final model performance. Specifically, a set of 181 TFs was extracted from the ENCODE database [6], and only those related to the gene in the 5p' position were considered. Additionally, all GOs involving fused genes were selected.

Finally, we included all miRNAs predicted to target all 5p' and 3p' genes to the feature set. This information was extracted from TargetScan, a popular state-of-the-art database that predicts biological targets of miRNAs by searching for the presence of sites that match the seed region of each miRNA [1], reporting for each miRNA all possible target genes. A set of 333 miRNAs was obtained by investigating the probability to target the genes belonging to the gene fusion. In case of ambiguity, only the highest probability was retained. To our knowledge, this is the first time that post-transcriptional regulation information has been used in such a classification task.

The final feature number was 5644, which is a considerably high number compared to the number of samples in our training and test sets. Thus, we performed feature selection to reduce feature set size to avoid overfitting our dataset. The chosen feature selection method was Random Forest, by which the number of features was lowered to 310.

3. Results

This section discusses the results obtained with ChimerDriver and the comparison with the state-of-the-art tools. Additionally, a case study in which ChimerDriver was applied on a pair of well-known datasets is presented.

3.1. *Results on the training set*

As previously stated, ChimerDriver was trained on 1765 gene fusions, obtained from COSMIC, Catalog of Somatic Mutations in Cancer [13] and from Babicenau et al. work [3]. Given each gene fusion's breakpoint, the aforementioned features are extracted and then fed to the MLP. The model was cross-validated on the training set with the k-fold method. K value was set equal to 10. The AUC, Accuracy, F1 score, precision and recall are reported in Table 1. The model reached an average f1-score of 0.98 on our training set with different combinations of learning rate and dropout values.

3.2. *Results on the test set*

The model was tested on 4877 gene fusions. 2623 oncogenic gene fusions were retrieved from the work of Gao et al.[17] and the remaining 2254 were gene fusions found in healthy tissues and reported by Babicenau et al. [3].We ensured that the test samples are entirely independent from the training samples. The model returned a 0.83 f1-score and 96% recall when tested on this set of gene fusions.

3.3. *miRNA impact on the classification performance*

The miRNA features were extracted from TargetScan [1], a popular database that maps gene-miRNA pairs providing various kinds of information. We mainly focused on the miRNA's probability of targeting the specific gene during post-transcriptional regulation. This value was extracted for both 5p' and 3p' genes and it is intended to represent the involvement of miRNAs in gene fusion processes. In Fig. 2 we highlight the impact of the miRNA features in the classification by displaying the confusion matrices including and excluding miRNAs from the evaluation. The impact of miRNAs is particularly evident when looking at the number of false-negative gene fusions, which is almost doubled when miRNAs are not considered. Including miRNAs in the classification task increases the recall value from 93% to 96%.

3.4. *Comparison with state of the art*

ChimerDriver performances were compared to those reported by three related works: Oncofuse [34], DEEPrior [29], and Pegasus [14]. To compare the results in the most unbiased way, the experimental conditions of the three tools were reproduced and ChimerDriver was applied. The results presented in this section are also summarized, in Supplementary Table 3, in the Supplementary Materials.

3.4.1. *Oncofuse*

To test the robustness of the proposed method, we extrapolated the training set and testing set used by Oncofuse [34]. Those samples were used to train and test our model and compare Oncofuse and ChimerDriver performances.

Oncofuse training samples were extracted from TICDB [37], a curated database that contains gene fusions found in tumor samples, and from a collection of fusion genes [16], and read-through transcripts [35] found in normal cells named NORM-RTH. Oncofuse's authors then built the oncogenic testing set by merging oncogenic gene fusions from CHIMERDB [25] and NGS, respectively oncogenic fusions predicted by gene fusion detection tools and fusions discovered and published in NGS

Table 1
Cross validation results with the k-fold method. The value of k was set equal to 10.

Learn rate	Dropout	AUC	Accuracy	F1	Precision	Recall
0.0001	0.0	0.981	0.978	0.981	0.981	0.981
0.0001	0.2	0.979	0.976	0.979	0.980	0.979
0.0001	0.4	0.981	0.978	0.981	0.981	0.981
0.001	0.0	0.980	0.976	0.980	0.980	0.980
0.001	0.2	0.976	0.972	0.975	0.968	0.984
0.001	0.4	0.977	0.974	0.977	0.980	0.975
0.01	0.0	0.982	0.979	0.982	0.986	0.979
0.01	0.2	0.982	0.979	0.982	0.983	0.982
0.01	0.4	0.980	0.976	0.980	0.983	0.978

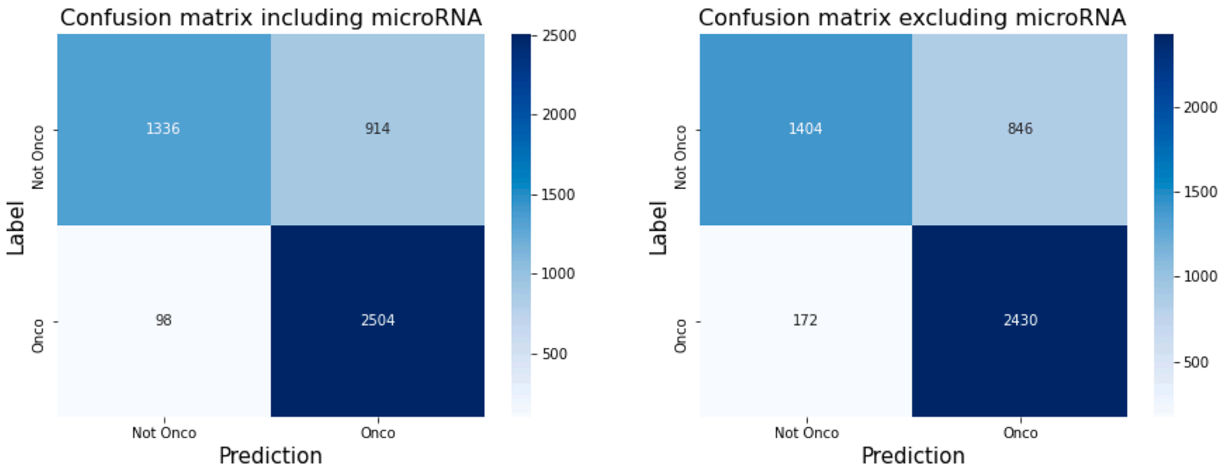


Fig. 2. Confusion matrices reporting the MLP results including miRNAs (on the left) and excluding miRNA features (on the right).

studies about cancer [2,5,9,41]. On the other hand, not-oncogenic testing samples were taken from Refseq [24] and CGC [43], two databases that report unbroken gene fusions. In particular, the samples that belong to CGC involve unbroken oncogenic genes.

All the previously listed features (see Material and methods for details) were processed and gathered, except for the two features related to the retained percentage of genes since the provided dataset omitted the breakpoint information.

The ChimerDriver model was tailored to this comparison. In detail, obtained 281 input features: the strands and the involvement in oncogenic processes of both 5p' and 3p' genes, 93 TFs, 155 miRNAs, and 30 GOs. The maximum number of epochs was set to 50, and the number of nodes per layer was 256, 128, 64, and 32 (the associated activation functions were the relu, sigmoid, relu, and sigmoid, respectively). The learning rate was fixed to 0.03, while the dropout value applied to each layer was 0.4.

Fig. 3 shows the comparison of the classification results obtained by ChimerDriver and Oncofuse. Precisely, the green bars correspond to the results achieved by Oncofuse, as reported by its authors [34], while the blue ones show the results obtained by ChimerDriver. Similar to Oncofuse paper, the results are displayed separately for each database. The bar diagram shows the percentage of driver gene fusions detected by the model. As it can be noticed, when trained and tested on the samples provided by Oncofuse, ChimerDriver provided better results with respect to those illustrated in the original paper.

Specifically, as reported by Fig. 3, 95% of TICDB samples were correctly classified as driver gene fusions by ChimerDriver as opposed to the assumed 90% reported by Oncofuse, furthermore 2% of the NORM-RTH samples were incorrectly classified as driver gene fusions by ChimerDriver as opposed to the assumed 10% reported by Oncofuse.

ChimerDriver successfully outperformed Oncofuse in the oncogenic gene fusion databases used as a test set, namely ChimerDB2A, ChimerDB2B, ChimerDB2C, NGS1 and NGS2. ChimerDriver identified more or a comparable amount of oncogenic gene fusions in each database with respect to Oncofuse, correctly classifying about 1/3 of the samples.

ChimerDriver minimized the number of detected driver fusions of unbroken oncogenic genes, identifying a lower number of driver gene fusions in CGC database, as additional test set.

When tested on the not-oncogenic samples in RefSeq database, Chimerdriver returned a slightly higher number of driver gene fusions.

In general, we may conclude that even without the information on the retained percentage of genes, ChimerDriver outperformed Oncofuse in the great majority of cases.

3.4.2. DEEPrior

DEEPrior is a DL-based classifier that performs gene prioritization using protein sequences obtained from the gene fusion samples. Its architecture is based on a CNN and an LSTM network. It was trained on a dataset extracted from COSMIC [13], and Babicenau et al.'s study [3] and tested on the part of the oncogenic gene fusion collection validated by Gao et al. [17]. DEEPrior reconstructs the protein sequences from gene fusion breakpoint information and assigns to each gene fusions an

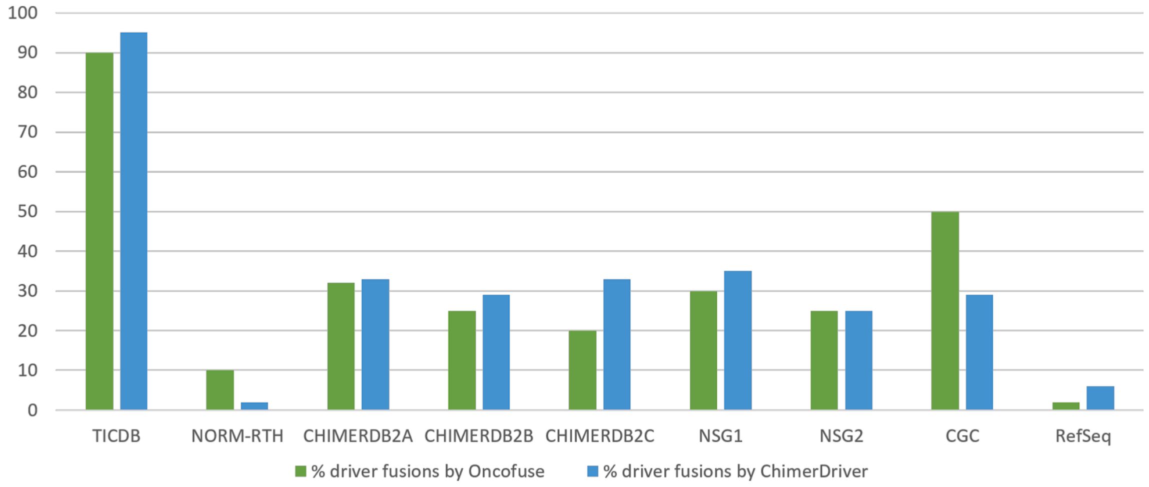


Fig. 3. The green bars correspond to the results reported by Shugay M. et al. in their paper. In blue the results obtained by ChimerDriver are displayed.

oncogenic score defining its oncogenic probability. Gene fusions are ordered according to the oncogenic score and highly scored fusions are prioritized as drivers. In this sense, DEEPrior main aim consists in providing a reliable classification prediction (oncogenic or not) according to the oncogenic score.

We trained and tested ChimerDriver on the DEEPrior training set and test set (*Dataset 2* in DEEPrior paper). As a result, ChimerDriver correctly classified 96% of oncogenic gene fusions from the test set. On the contrary, DEEPrior prioritized as driver the 32.48% of gene fusions found in the test set. Since DEEPrior aims at classifying only highly probable oncogenic fusions, the percentage of prioritized gene fusions is not directly comparable with the classification performances obtained with ChimerDriver. ChimerDriver provides a classification result for each gene fusion, while DEEPrior classifies a tiny percentage of gene fusions in the dataset.

We can conclude that the ChimerDriver approach exploits different sources of information (TFs, GOs, miRNAs) while DEEPrior focuses on identifying the oncogenic potential of a gene fusion through its protein sequence without considering the effect of post-transcriptional regulators.

At the same time, ChimerDriver ensures a less computationally intensive approach in the training phase than DEEPrior.

3.4.3. *Pegasus*

To further assess ChimerDriver classification performances, we took into account Pegasus [14], a state-of-the-art tool for gene fusion detection and classification purposes. Pegasus exploits a traditional machine learning model to predict of driver gene fusion, namely a gradient tree boosting algorithm.

Also, in this case, ChimerDriver was trained and tested on the gene fusion samples used to develop and validate Pegasus.

We observed that the training dataset was strongly unbalanced towards the negative samples, comprising over 9923 negative samples out of 10162 gene fusions. Not to penalize the MLP architecture, which is particularly sensitive to class unbalance, we lowered the number of negative gene fusions to 239, namely the number of positive samples.

ChimerDriver was cross-validated on 10 folds using the aforementioned training samples. It should be noted that, as a result of balancing the classes, the model was given a fairly small number of training examples. In the end, the f1-score was equal to 0.89 with a learning rate and dropout, respectively equal to 0.001 and 0.

Pegasus' test set accounted for 78 gene fusions, 39 oncogenic and 39 not oncogenic, respectively. According to Pegasus authors, the curated subset of 39 oncogenic gene fusions was almost entirely correctly classified by Pegasus, which reported 0.97 of AUC and 0.95 of AUC for the not oncogenic samples.

Pegasus intently selected as negative examples 39 not oncogenic gene fusions containing at least a tumor suppressor or an oncogene. The rationale is that these gene fusions would be most challenging for a classification task. ChimerDriver correctly classified 27 out of the 39 not-oncogenic gene fusions enforcing the notion that the model can generalize even on not oncogenic gene fusions. On the other hand, the oncogenic test samples represented a more difficult classification task for ChimerDriver, which detected 17 oncogenic gene fusions. It should be noted that ChimerDriver model was initially trained and tested on a wide variety of gene fusions proving its ability to learn and generalize well when given a fair amount of examples. On the contrary, since Pegasus was developed and refined on particular tissues, a reduced number of samples is used as a training set.

In our opinion, the small number of samples in the Pegasus training set negatively impacted the ChimerDriver training phase, which benefits from a wider number of gene fusions. Therefore, ChimerDriver performances, when trained and tested on Pegasus datasets, are negatively affected, hindering the likelihood of reaching the outcome reported by the Pegasus authors.

3.5. *Case study*

Finally, to assess ChimerDriver's performances in a clinical context, we selected two well-known studies: 6 breast cancer samples [8] and 4 prostate cancer samples [46] in which 24 gene fusions are reported to be experimentally validated. The samples are all RNA-seq data. We processed them with STAR-fusion [19] to identify which gene fusions were found in these samples by a standard and accurate fusion detection tool. 21 out of the 24 validated gene fusions were detected with STAR-fusion and subsequently processed with ChimerDriver to confirm the ability to detect oncogenic gene fusions in a real-world case correctly. Fig. 4 shows the results of this assessment. Specifically, the gene fusions marked in gray were not detected by STAR-fusion hence were not available to ChimerDriver for further processing. The training dataset and the training parameters are described in detail in the Material and methods section like the ones generally used in the ChimerDriver training procedure. On the 21 samples, ChimerDriver wrongly classified as not oncogenic the three oncogenic gene fusions marked in orange. By inspecting the oncogenic role of 5p' and 3p' genes and the retained percentage in the gene fusion, a possible explanation for the wrong classification could be hypothesized. Concerning the ACACA-STAC2 gene fusion, no information on the involvement of any of the two genes was provided to the algorithm. So, although most of the portion of both genes was retained after the gene fusion event, ChimerDriver was probably unsure about their role in oncogenic processes. As for the GLB1-CMTM7 fusion, the algorithm was aware that the latter gene is involved in tumor suppression; on the other hand, the retained percentage of CMTM7 was less than 45%. This probably led the network to conclude that there was not enough gene left in the gene fusion to cause issues. Similarly, in the CPNE1-PI3 fusion, the percentage of retained genes (respectively 25% and 40%) was probably too low to label the gene fusion as oncogenic even if the genes were associated with the roles oncogenic and driver, respectively. Finally, ChimerDriver correctly classified the 18 remaining gene fusions as oncogenic. Hence, ChimerDriver correctly classified 18 out of 21 oncogenic gene fusions, demonstrating that the specifically designed neural network is proficient in learning and generalizing from a consistent number of gene fusion samples. Moreover, the information gathered from the different sources and provided to the tool as features proved to be particularly effective in discerning oncogenic and not-oncogenic fusions even in a realistic circumstance.

4. Discussion

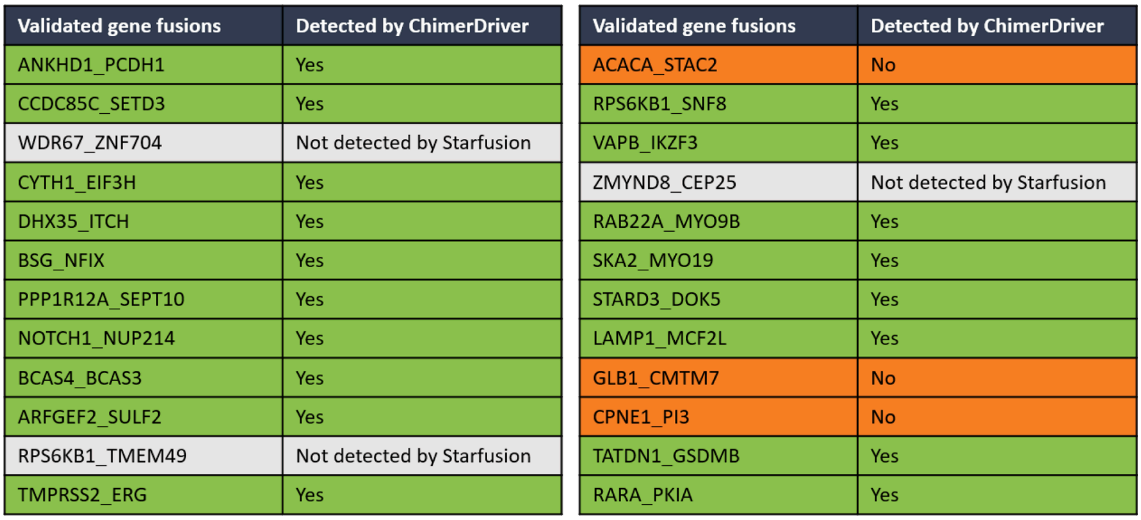
Identifying oncogenic gene fusions is of crucial importance in cancer detection and prognosis. To date, state-of-the-art tools exploit transcriptional and GOs information without considering the post-transcriptional regulators in predicting the oncogenic potential of a gene fusion. Here, we presented ChimerDriver, a novel tool to accomplish the aforementioned task exploiting transcriptional and post-transcriptional regulators. In detail, ChimerDriver focuses on miRNAs post-transcriptional effect as a key feature to perform the prediction.

ChimerDriver is based on an ad-hoc designed neural network embedding miRNAs, transcription factors, gene ontologies, and gene-specific information to predict gene fusions' oncogenic potential. The model is stable and exhibits excellent classification performance (f1-score = 0.98).

We tested our classifier against three state-of-the-art tools: Oncofuse, DEEPrior, and Pegasus.

With respect to Oncofuse, we introduced post-transcriptional regulation to perform the classification and, as a result, ChimerDriver outperformed Oncofuse in the great majority of tested cases.

In particular, ChimerDriver performed better than Oncofuse on the test set, correctly classifying as oncogenic about 1/3 of the oncogenic gene fusions. ChimerDriver identified a comparable or higher amount of oncogenic gene fusions outperforming Oncofuse results in each positive



Validated gene fusions	Detected by ChimerDriver
ANKHD1_PCDH1	Yes
CCDC85C_SETD3	Yes
WDR67_ZNF704	Not detected by Starfusion
CYTH1_EIF3H	Yes
DHX35_ITCH	Yes
BSG_NFIX	Yes
PPP1R12A_SEPT10	Yes
NOTCH1_NUP214	Yes
BCAS4_BCAS3	Yes
ARFGEF2_SULF2	Yes
RPS6KB1_TMEM49	Not detected by Starfusion
TMPRSS2_ERG	Yes

Validated gene fusions	Detected by ChimerDriver
ACACA_STAC2	No
RPS6KB1_SNF8	Yes
VAPB_IKZF3	Yes
ZMYND8_CEP25	Not detected by Starfusion
RAB22A_MYO9B	Yes
SKA2_MYO19	Yes
STARD3_DOK5	Yes
LAMP1_MCF2L	Yes
GLB1_CMTM7	No
CPNE1_PI3	No
TATDN1_GSDMB	Yes
RARA_PKIA	Yes

Fig. 4. The 24 oncogenic gene fusions validated in prostate and breast tumor samples are reported. STAR-fusion did not detect the three gene fusions marked in gray hence were not available to ChimerDriver for further processing. ChimerDriver correctly classified as oncogenic 18 out of the 21 available gene fusions.

test case. ChimerDriver minimized the number of detected driver fusions in ’unbroken oncogenic genes’ (negative testing samples) extracted from CGC compared to Oncofuse. This result confirmed the ability of ChimerDriver in generalizing and taking advantage of the given set of features to make a correct prediction. As previously presented in the Results section about Pegasus comparison, this statement is true even when the samples contain an oncogene or a tumor suppressor. ChimerDriver returned a slightly higher number of oncogenic gene fusions than Oncofuse when tested on the RefSeq database of ’unbroken not-oncogenic genes’. We recall that the breakpoint information was not available in Oncofuse datasets. Therefore, to perform an unbiased comparison with Oncofuse, the breakpoint information was neglected by the ChimerDriver model. Consequently, the percentage of driver gene fusions detected by ChimerDriver on RefSeq was slightly higher than expected, probably because the tool could not profit from the breakpoint information.

ChimerDriver also outperformed DEEPrior in terms of the number of classified gene fusion. In particular, ChimerDriver correctly identified 96% of oncogenic gene fusions in the dataset used to test DEEPrior, which prioritized as oncogenic only 32.48% of the samples. It should be noted that the goals of DEEPrior and ChimerDriver are slightly different. The first prioritizes gene fusions, returning those with an oncogenic probability greater than a threshold (typically 80%). ChimerDriver performs an immediate classification of each gene fusion by integrating transcriptional and post-transcriptional features in the assessment. The outcome of ChimerDriver is remarkable in terms of the number of oncogenic samples that were correctly classified while also enlightening because it stresses the extent to which miRNAs are involved in the oncogenic processes of gene fusions.

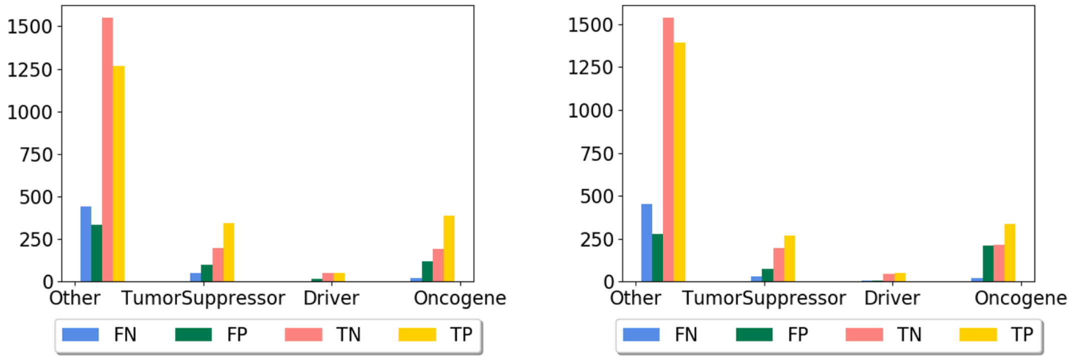
Moreover, the performances of ChimerDriver were compared to the ones reported by Pegasus authors. According to their research, the latter could correctly classify almost all of the test samples. After training and testing ChimerDriver on the gene fusions provided by the authors, it was observed that the number of detected oncogenic samples was lower than the results reported by Pegasus. As already stated in the Results section, the number of training samples was lowered in order to balance the oncogenic and not oncogenic classes. However, the limited number of samples processed by ChimerDriver in the training phase has probably inhibited the neural network from learning efficiently. In addition, Pegasus’s authors specify that the negative validation samples included at least one oncogene or tumor suppressor. We remind that, to make a prediction, ChimerDriver also relies on the role of each gene in oncogenic processes (e.g., driver, oncogene, or tumor suppressor), making the classification task particularly arduous to tackle. In addition, Pegasus and, consequently, ChimerDriver were trained on a reduced number of samples, thus impacting ChimerDriver performances. Nevertheless, ChimerDriver correctly classified most of the not oncogenic gene fusions enforcing the notion that the model can generalize well in this situation.

In this work, we focused on the integration of information coming from different databases to improve the current state-of-the-art research on classifying oncogenic gene fusions. Additionally, a neural network was designed explicitly for this task. However, the main contribution of the present work is the introduction of miRNAs in the classification model. In fact, despite miRNAs role in determining the oncogenic potential of gene fusions has been demonstrated, they had never been considered in such a task. In the present work, we showed that they could significantly improve the model performance. In particular, they halved the number of false negatives and improved the recall of the model. We can conclude that miRNAs, being involved in the regulation of gene fusion-related protein, are a promising indicator of the oncogenic potential of gene fusions.

The main limitation of the proposed method is that some gene fusions are misclassified. To better investigate ChimerDriver classification with respect to the Cancermine [22] role, we reported in Fig. 5 the distribution of the Cancermine roles (e.g. tumor suppressor, driver, oncogene, other) for 5p’ gene (Fig. 5a) and 3p’ gene (Fig. 5b). In addition, test set samples are divided in each role according to the classification results (false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN)). TP samples are characterized for 5p’ ad 3p’ genes by a prevalence of suppressors and oncogenes. On the contrary, TN mostly refer to the ’other’ CancerMine role. Consequently, FP samples could consist of oncogenes (in particular for 3p’ gene) and FN samples are hardly ever related to tumor suppressors, drivers, or oncogenes. In this sense, FP and FN samples reflect ChimerDriver behavior on TP and TN, respectively. In a clinical context, FN misclassified samples are unlikely to be tested for in-lab validation since most involve genes with no specific oncogene/tumor suppressor role. FP samples instead would have been considered for experimental validation, which would exclude them from oncogenic fusions. However, laboratories would still benefit from a selection of putative oncogenic gene fusions.

5. Conclusions

Gene fusions are a common mutation that is nowadays known to be responsible for about 1/5 of human cancers. It is of the uttermost



(a) (b)

Fig. 5. Here we report the distribution of the false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) regarding Cancermine information for both 5p' and 3p' genes (respectively Fig. 5(a) and 5(b)). Noticeably, FPs are never tumor suppressors, drivers or oncogenes.

importance to correctly identify gene fusions to improve cancer detection and prognosis. Considering that the state-of-the-art tools exploit transcriptional and gene information neglecting post-transcriptional regulations, we combined this knowledge and established the value of miRNAs in achieving superior classification performances.

To conclude, we presented ChimerDriver, a novel and stable DL architecture based on a Multi-Layer Perceptron (MLP), that, for the first time, combines gene-level features with TFs and miRNAs targetting the gene fusion to perform its classification and prioritization.

ChimerDriver was trained and tested on a consistent number of gene fusions. The final results highlight the impact of miRNAs in evaluating the oncogenic potential of gene fusions. We can infer that the inclusion of miRNAs represents a valuable advantage in identifying oncogenic gene fusions.

ChimerDriver can become a valuable tool for research laboratories to predict the oncogenic potential of gene fusions. Indeed, the expensive validations could be targeted cost-effectively with this easy-to-use tool; additionally, it may speed up identifying novel and potentially oncogenic gene fusions, allowing for better diagnosis, classification, and treatment of cancer patients.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work has been supported by the SmartData@PoliTO center on Big Data and Data Science.

Funding

This study was funded by the European Union's Horizon 2020 research and innovation programme DECIDER under Grant Agreement 965193.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2022.104057.

References

[1] Vikram Agarwal et al., Predicting effective microRNA target sites in mammalian mRNAs, in: Elisa Izaurralde (Ed.), eLife 4 (2015) e05005. https://doi.org/10.7554/eLife.05005. issn: 2050–084X.

[2] Yan W. Asmann, et al., Detection of redundant fusion transcripts as biomark- ers or disease-specific therapeutic targets in breast cancer, Cancer Res. 72 (8) (2012) 1921–1928.

[3] Mihaela Babiceanu, et al., Recurrent chimeric fusion RNAs in non-cancer tissues and cells, Nucl. Acids Res. 44 (6) (2016) 2859–2872.

[4] Pietro Barbiero, et al., Unsupervised Multi-omic Data Fusion: The Neu- ral Graph Learning Network, in: International Conference on Intelligent Computing, Springer, 2020, pp. 172–182.

[5] Matteo Benelli, et al., Discovering chimeric transcripts in paired-end RNA- seq data by using EricScript, Bioinformatics 28 (24) (2012) 3232–3239.

[6] ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project, Science 306(5696) (2004) 636–640. https://doi.org/10.1126/science.1105136.

[7] Brian J. Druker, Imatinib as a paradigm of targeted therapies, Adv. Cancer Res. 91 (1) (2004) 1–30.

[8] Henrik Edgren, et al., Identification of fusion genes in breast cancer by paired-end RNA-sequencing, Genome Biol. 12 (1) (2011) 1–13.

[9] Henrik Edgren et al., Inga Rye, Sandra Nyberg, Maija Wolf, Anne Lise Borresen Dale et Olli Kallioniemi: Identification of fusion genes in breast cancer by paired-end RNA-sequencing, Genome Biol. 12.1 (2011) R6.

[10] E. Heyer Erin et al., Diagnosis of fusion genes using targeted RNA se- quencing, Nat. Commun. (2019). https://doi.org/10.1038/s41467-019-09374-9.

[11] Mitelman Felix, Johansson Bertil, Mertens Fredrik, The impact of translocations and gene fusions on cancer causation, Nat. Rev. Cancer 7 (2007) 233–245.

[12] Witold Filipowicz, Suvendra N. Bhattacharyya, Nahum Sonenberg, Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. 9 (2) (2008) 102–114.

[13] Simon A. Forbes, et al., COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer, Nucl. Acids Res. 39 (suppl 1) (2010) D945–D950.

[14] Abate Francesco et al., Pegasus: A Comprehensive Annotation and Pre- diction Tool for Detection of Driver Gene Fusions in Cancer, BMC Syst. Biol. (2014). https://doi.org/10.1186/s12918-014-0097-z.

[15] Milana Frenkel-Morgenstern et al., Chimeras taking shape: potential func- tions of proteins encoded by chimeric RNA transcripts, Genome Res. 22(7) (2012) 1231–1242.

[16] Milana Frenkel-Morgenstern, et al., Chimeras taking shape: potential func- tions of proteins encoded by chimeric RNA transcripts, Genome Res. 22 (7) (2012) 1231–1242.

[17] Qingsong Gao, et al., Driver Fusions and Their Implications in the De- velopment and Treatment of Human Cancers, Cell Rep. (2018), https://doi.org/10.1016/j.celrep.2018.03.050.

[18] Matt W Gardner, S.R. Dorling, Artificial neural networks (the multi- layer perceptron)–a review of applications in the atmospheric sciences, Atmos. Environ. 32 (14-15) (1998) 2627–2636.

[19] Brian J. Haas, et al., Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods, Genome Biol. 20 (1) (2019) 1–16.

[20] Brian J. Haas et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq, bioRxiv (2017). https://doi.org/10.1101/120295. eprint: https://www.biorxiv.org/content/early/2017/03/24/120295.full.pdf.url:https://www.biorxiv.org/content/early/2017/03/24/120295.

[21] Matthew K. Iyer, Arul M. Chinnaiyan, Christopher A. Maher, ChimeraS- can: a tool for identifying chimeric transcription in sequencing data, Bioinformatics 27 (20) (2011) 2903–2904.

[22] Lever Jake, et al., CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer, Nat. Methods 16 (2019) 505–507, https://doi.org/10.1038/s41592-019-0422-y.

[23] Kalpana Kannan et al., Recurrent BCAM-AKT2 fusion gene leads to a constitutively activated AKT2 fusion kinase in high-grade serous ovarian carcinoma, Proc. Natl. Acad. Sci. 112(11) (2015) E1272–E1277. https://doi.org/10.1073/pnas.1501

735112. issn: 0027-8424. eprint: https://www.pnas.org/content/112/11/E1272. full.pdf. url: https://www.pnas.org/content/112/11/E1272.

[24] W. James Kent, et al., The human genome browser at UCSC, Genome Res. 12 (6) (2002) 996–1006.

[25] Pora Kim, et al., ChimerDB 2.0–a knowledgebase for fusion genes up- dated, Nucleic Acids Res. 38 (suppl 1) (2010) D81–D85.

[26] M. Natividad Lobato, et al., Modeling chromosomal translocations using conditional alleles to recapitulate initiating events in human leukemias, J. Natl. Cancer Inst. Monogr. 2008 (39) (2008) 58–63.

[27] Marta Lovino et al., A deep learning approach to the screening of onco- genic gene fusions in humans, Int. J. Mol. Sci. 20(7) (2019) 1645.

[28] Marta Lovino, et al., A survey on data integration for multi-omics sample clustering, Neurocomputing (2021).

[29] Marta Lovino et al., DEEPrior: a deep learning tool for the prioritiza- tion of gene fusions, Bioinformatics 36(10) (2020) 3248–3250. https://doi.org/10.1093/bioinformatics/btaa069. issn: 1367-4803. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/10/3248/33204199/btaa069.pdf.url:https://doi.org/10.1093/bioinformatics/btaa069.

[30] Marta Lovino, et al., Multi-omics Classification on Kidney Samples Ex- ploiting Uncertainty-Aware Models, in: International Conference on In- telligent Computing, Springer, 2020, pp. 32–42.

[31] Natalia J. Martinez, Albertha J.M. Walhout, The interplay between transcription factors and microRNAs in genome-scale regulatory networks, Bioessays 31 (4) (2009) 435–445.

[32] Andrew McPherson, et al., deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data, PLoS Comput. Biol. 7 (5) (2011) e1001138.

[33] Fredrik Mertens, Cristina R. Antonescu, Felix Mitelman, Gene fusions in soft tissue tumors: recurrent and overlapping pathogenetic themes, Genes Chromosomes Cancer 55 (4) (2016) 291–310.

[34] Mikhail Shugay, Iñnigo Ortiz de Mendíibil, José L. Vizmanos, Francisco J. Novo, Oncofuse: A Computational Framework for the Prediction of the Oncogenic Potential of Gene Fusions, Bioinformatics 29 (20) (2013) 2539–2546. https://doi.org/10.1093/bioinformatics/btt445.

[35] Serban Nacu, et al., Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples, BMC Med. Genomics 4 (1) (2011) 1–22.

[36] Mridula Nambiar, Vijayalakshmi Kari, Sathees C. Raghavan. Chromosomal translocations in cancer, Biochim. Biophys. Acta (BBA) Rev. Cancer 1786(2) (2008) 139–152.

[37] Francisco J. Novo, Inigo Ortiz de Menibil, José L. Vizmanos, TICdb: a collection of gene-mapped translocation breakpoints in cancer, BMC Genomics 8(1) (2007) 1–5.

[38] Sankar K. Pal, Sushmita Mitra, Multilayer perceptron, fuzzy sets, classifiaction (1992).

[39] Ilaria Roberti, et al., Exploiting Gene Expression Profiles for the Auto- mated Prediction of Connectivity between Brain Regions, Int. J. Mol. Sci. 20 (8) (2019) 2035.

[40] Dennis W. Ruck et al., The multilayer perceptron as an approximation to a Bayes optimal discriminant function, IEEE Trans. Neural Netw. 1(4) (1990) 296–298.

[41] Onur Sakarya, et al., RNA-Seq mapping and detection of gene fusions with a suffix array algorithm, PLoS Comput. Biol. 8 (4) (2012) e1002464.

[42] Alice T. Shaw, et al., Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis, Lancet Oncol. 12 (11) (2011) 1004–1012.

[43] Zbyslaw Sondka, et al., The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers, Nat. Rev. Cancer 18 (11) (2018) 696–705.

[44] Nicolas Stransky, et al., The landscape of kinase fusions in cancer, Nat. Commun. 5 (1) (2014) 1–10.

[45] Scott A. Tomlins et al., Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer, Neoplasia 10(2) (2008) 177–IN9., https://doi.org/10.1593/neo.07822.url:http://www.sciencedirect.com/science/article/pii/S1476558608800644. issn: 1476-5586.

[46] Wu. Chunxiao, et al., Poly-gene fusion transcripts and chromothripsis in prostate cancer, Genes Chromosomes Cancer 51 (12) (2012) 1144–1153.

[47] Andrew D. Yates, Premanand Achuthan, Akanni, et al.. Ensembl 2020, Nucleic Acids Res. 48(D1) (2019) D682–D688. doi: https://doi.org/10.1093/nar/gkz966. eprint: https://academic.oup.com/nar/article-pdf/48/D1/D682/31697830/gkz966.pdf.url:https://doi.org/10.1093/nar/gkz966. issn: 0305–1048.