Transcription factor binding site detection using convolutional neural networks with a functional group-based data representation

Gergely Pap[1], Györgypál Zoltán[2], Krisztián Ádám[2], László Tóth[1], Zoltán Hegedűs[2,3]

[1] Institute of Informatics, University of Szeged, Árpád Square 2, H-6720 Szeged, Hungary
[2] Institute of Biophysics, Biological Research Centre, H-6726 Szeged, Temesvári Blvd. 62 Hungary
[3] Department of Biochemistry and Medical Chemistry, University of Pécs, Pécs, Hungary

{papg,tothl}@inf.u-szeged.hu, hegedus@brc.hu

Abstract. Transcription factors (TFs) play an essential role in molecular biology by regulating gene expression. The binding sites of TFs can vary by a large amount and the numerous possible binding locations make their detection a challenging issue. Recently, several machine learning approaches using nucleotide sequence data were applied to classify DNA sequences regarding Transcription Factor Binding Sites (TFBS). We propose a novel training strategy without the traditional 1D nucleotide-based DNA sequence representation by instead using a 2D topological matrix of sub-nucleotide chemical functional groups substantially defining the protein binding ability of DNA fragments. We train convolutional neural networks using this novel Functional Group DNA Representation (FGDR) to solve a TFBS classification task. We compare our results with the efficiency of previous nucleotide-based training approaches and show that learning from an FGDR data sequence has several benefits regarding TFBS classification. Moreover, we reason that learning deep neural networks from the FGDR representation produces competitive results while only introducing a pre-processing conversion step. Finally, we show that employing an ensemble of models from the nucleotide and FGDR representations for network training results in higher classification performance than any of the single input approaches.

1. Introduction

Transcription factors (TFs) are gene expression regulating proteins which play an important role in almost all cell physiological processes and in the related molecular mechanisms. Transcription factors detect and bind DNA double helix strands at TF specific positions called DNA recognition motifs. Motifs are represented by the sequential combination of A-C-G-T nucleotides and are typically 4-18 base-pair long. Finding and classifying these motifs is a long-standing question of molecular and computational biology. Next Generation Sequencing (NGS) provides an ample amount of raw nucleotide sequence data which needs to be processed and analysed in order to achieve further research goals. While experimental identification of TFBS is an expensive and time-consuming process, efficient in silico methods predicting specific binding sites for the investigated TF can

substantially facilitate the process. One of the most promising recent approaches for in silico TFBS prediction is the application of deep learning frameworks since deep learning-based applications are excellent tools for recognising patterns in large amounts of data. Moreover, in the last few years previous bioinformatics methods based on position weight matrices and other interpretable statistical methods for identification of DNA recognition motifs were surpassed by machine learning approaches trained on nucleotide sequence data. On the other hand, the conventional nucleotide-based DNA representation describes the real contact forming chemical sub-structure pattern only in an implicit manner. We believe the explicit representation of this information in the training data can be beneficial for the learner as it more precisely describes the underlying molecular mechanisms. Learning CNNs on this novel representation for TFBS classification surpass the performance of other, nucleotide sequence-based methods.

2. Related work

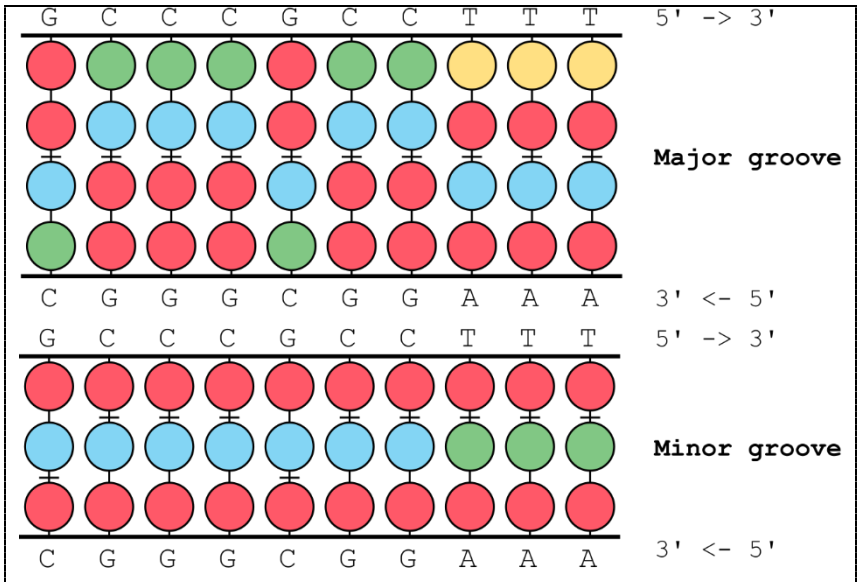
The ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge data set provided a suitable benchmark and Artificial Intelligence (AI) test setting for TFBS prediction. DeepBind [1] produced the highest classification accuracies in 2015 and was the gold standard for this particular task. Later Zeng et al. [2] examined the DeepBind-based CNNs architecture (which is an exceptionally important property of the machine learning approach) and published several useful and effective empirical results. Others also continued and improved upon the ideas presented in the DeepBind article. iDeep [3] uses sequence data with CNNs and numerical features with Deep Belief Networks so that the networks complement each other and produce more efficient predictions, which are the benefit of the multi-modal input method. DeeperBind [4] leverages the advantage of recurrent neural networks by using Long Short-Term Memory (LSTM) cells after the convolutional approach based on nucleotide data sequence. Fu et al. [5] introduced scFAN (Single Cell Factor Analysis), a system that takes advantage of a pre-trained network on bulk-data which, after fine-tuning, makes predictions with single-cell information. FactorNet [6] uses LSTM cells after the convolutional blocks produced the feature maps. Wenxiu et al. [7] proposed SVMs fitted on DNA conformation and sequence data, where the shape descriptors of DNA would include conformational properties of the DNA stands and the base-pairs forming them, (e.g. MGW, Roll, ProT and HelT) and achieved better performance with the combination of sequence and shape features than those based solely on DNA sequence. The previously mentioned studies were mostly based on nucleotide sequence training data. Also, the use of CNNs to capture the features' local dependencies (and in some cases, RNNs to account for temporal information) were successfully accompanied and complemented by other forms of input data (apart from nucleotide sequence). To our knowledge, this article is the first machine learning experiment using a more detailed chemical DNA description, based on the surface-accessible functional groups of the double helix (Functional Group DNA Representation, FGDR).

3. Functional Group DNA Representation

Previous studies interested in machine learning with bioinformatical tasks and nucleotide sequence data in the majority of cases convert the 4 nucleotides of DNA to a One-Hot Encoding scheme, where a sequence of length L is shaped into a 4 by L matrix and each row corresponds to a nucleotide by having a value of 1 if the sequence contained that specific nucleotide at the given position and 0 otherwise. A more detailed description of FGDR used for the training of CNNs can be found in the DRV [8] article. However, instead of nucleotide sequences, our approach uses a functional group-based one, an explicit chemical description of the given DNA fragment. The motivation behind using FGDR for TFBS prediction is that TF bindings to DNA strands are dependent upon the chemical structure of the binding sites' nucleotides. Using the ACGT description as training input might not be optimal because relevant information about the TFBS is coded implicitly and we argue that explicitly making the functional group information available to the network during training brings potential benefits and this additional information increases performance (see Experiments and Results, Table 2.).

4. Data set and representation description

We decided to test our approach using the data from the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge since many other works were also conducted on this benchmark data set. We acquired the sequences from http://cnn.csail.mit.edu/ [2]. The data set of TFBS uses 101 nucleotides and the number of training entities (sequences of ACGT) belonging to a TF is (roughly) between 10k and 100k. For our studies 10 TF data sets were chosen from the 'motif discovery' task of the ENCODE-DREAM challenge for experimenting based on biological and machine learning standpoints such as: TF family, binding mechanism, cell line, antibody type, number of training sequences and prediction accuracy achieved by previous works. The nucleotide-based DNA sequences were converted to FGDR format using a custom R script (for further information, refer to the DRV site [8]). The FGDR representation by default is a 7 x L matrix where L is the sequence length and the 7 rows are representing the topological positions of the different chemical functional groups within the major and minor grooves of the DNA double helix. 4 rows of the FGDR matrix come from the major and 3 from the minor groove. The chemical functional groups are represented as categorical values reflecting the donor and acceptor capabilities of the functional groups during the H-bond formation process. The FGDR scale uses values ranging from 0 to 8: CNNs receive FGDR input in a numerical form which is approximately proportional with the donor-acceptor property of the functional groups (i.e., 0: strong donor; 3 and 4: weak donor, 8: strong acceptor) as can be seen in Fig 1. We normalized the values in the 7 x L FGDR matrix to 0-1.



8	4	4	4	8	4	4	3	3	3
8	0	0	0	8	0	0	8	8	8
0	8	8	8	0	8	8	0	0	0
4	8	8	8	4	8	8	8	8	8
8	8	8	8	8	8	8	8	8	8
0	0	0	0	0	0	0	4	4	4
8	8	8	8	8	8	8	8	8	8

Figure 1 Top: Conversion of the sequence [GCCCGCCTTT] to FGDR. Bottom: The functional group categories are assigned a numerical value representing them as a donor: 0, methyl: 3, hydrogen 4 or acceptor: 8.

5. Experiments and Results

5.1. Training setup

For training and evaluation 10 TFs from the 'motif discovery' task were used. The classification results shown in this study were all computed from the respective TF's given test set. To account for the stochastic nature of deep learning, each experiment was repeated 5 times and the results were averaged. 10% of the training data was separated in order to construct a validation set to configure the hyperparameters and to monitor performance during training.

5.2. Network structure

The Deepbind network uses 1 layer of convolution and the work of Zeng et al. searches for an optimal architecture setup resulting in relatively shallow networks (1-3 convolutional layers) compared to the often deep, 100+ convolutional layers of networks applied in computer vision or image processing. Since FGDR is a larger input space compared to nucleotide data, we found that constructing an adequately complex or deep CNN is necessary for accurate model performance. We experimented with shallower and deeper architectures than the one described in the following but found them respectively underfitting or overfitting. In parallel with nucleotide-based TFBS classifier CNNs mentioned in Section 2 [4, 2], we also define the channels of the CNN's input layer as the rows of the FGDR matrix so that each channel corresponds to one of the 7 vectors of the input matrix containing the FGDR chemical properties. We decided to start with a neural network architecture similar to that of DeepBind, so the input data is fed to quasi 1D convolutional layers to construct the feature maps. In our implementation we used 2D convolutions with one of the dimensions set to 1 and our input matrix shape also had an extra axis of shape 1.

5.3. Hyperparameter optimization

It is well-established in deep learning literature that network performance is highly dependent on the proper selection of hyperparameters. Because this paper's original goal is to show the potential in FGDR-based learning, for our initial training setup we followed the structures defined in [1, 2]. However, we found that the novel (and possibly more complex) input type requires deeper networks and more thorough considerations regarding architecture, regularization and the optimizer. Deviating from the networks' complexity of previous works might blur the difference about data representation benefits, but considering the novel input format, some structural and parameter changes were necessary to facilitate well-trained models. The hyperparameters of the learners were established empirically by grid-search. Two convolutional layers proved to be optimal with 768 filters each. For the kernel size, an appropriate setting was 1x24 in the first and 1x12 in the second layer of convolution. The convolutional layers used linear activation and a ReLU activation was applied after finishing the computations involved in the convolutions' regularization. To avoid overfitting, a L2 kernel- and an L1 activity regularizer were introduced with values of 5e-4 and 5e-5, respectively. A maximum pooling operation was applied after the second convolution layer to get the extracted features in a reduced size from the convolutional blocks. Then two fully connected layers with dropout (500 and 150 units and a dropout probability of 0.5 in the larger layer) were next and finally a softmax output layer for classification followed. The ADAM optimizer was selected with an initial learning rate of 5e-5 and a learning rate decay was applied monitoring the changes on the validation set's loss values.

Table 1. Accuracy and F1 values for 10 different Transcription Factors. The best results are produced by averaging the probabilities of the Nuc. and the FGDR networks (denoted as Ens.).

TF	Acc. Nuc.	Acc. FGDR	Acc. Ens.	F1 Nuc.	F1 FGDR	F1 Ens.
SydhK562Znf143Iggrab	0.898	0.898	0.901	0.899	0.902	0.898
HaibH1hescSp1Pcr1x	0.787	0.791	0.794	0.713	0.711	0.713
SydhK562CjunIfng30	0.87	0.867	0.871	0.873	0.873	0.875
SydhK562CmycIggrab	0.796	0.8	0.804	0.795	0.802	0.805
SydhK562MaxIggrab	0.808	0.815	0.818	0.816	0.816	0.821
SydhK562Mxi1af4185Iggrab	0.758	0.77	0.769	0.762	0.775	0.774
SydhImr90MafkIggrab	0.943	0.942	0.944	0.943	0.942	0.945
SydhGm12878Mazab85725	0.771	0.772	0.779	0.78	0.771	0.782
HaibK562Yy1V0416102	0.838	0.838	0.842	0.751	0.746	0.754
UtaK562Ctcf	0.958	0.958	0.96	0.958	0.957	0.96

5.4. Comparing the nucleotide and FGDR networks

We selected 10 transcription factors from the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge. The data sets are available at http://cnn.csail.mit.edu/ [2]. The FGDR format uses more disk space and the conversion's time cost is considerable, while training a model using FGDR generally takes 5-10 minutes using modern NVIDIA GPUs. We shuffled the data before every training run then separated 10% of the train set for validation. We used the previous as our metric for early stopping. The test sets were always the same as provided in [2]. We ran 5-5 training runs for the nucleotide and FGDR data sets and measured their classification performance. We present the result of training on 10 TFs separately in Table 1., which provides ample evidence and helps keep the paper brief. In the cases of Znf143, Yy1 and Ctcf, the performance of the nucleotide and FGDR models match. The FGDR model has higher accuracy when classifying the Sp1, Cmyc, Max, Mxi and Maz TFs. The ensemble technique produces higher accuracies in the case of all TFs, except for Mxi, in which scenario the FGDR model performs better by 0.01.

Table 2. Comparing our methods to DeepBind and Zeng-CNN based on AUC. The measures were calculated by averaging the published AUC scores for the 10 TFs used in this paper.

Method	DeepBind	Zeng	Nuc.	FGDR	Ens.
AUC	0.863472	0.904524	0.9142	0.9145	0.9171

The FGDR models manage to surpass the nucleotide ones by a small margin when measuring AUC or accuracy. However, in terms of both metrics the ensemble approach produces the highest classification results. When two models with different inputs are used for predicting the class labels of the test set's entities and the resulting probabilities are averaged, the mean performance of classification accuracy increases to 84.8%. For comparison with other studies using the same benchmark data set, we provide the results presented in Table 2. Our models surpass the other works' performance based on AUC, though we note that our networks are more complex, have more trainable parameters and therefore a stronger representative capacity. This increase in complexity was necessary to produce models with good performance on FGDR, since the FGDR input is considerably larger than the one-hot encoded nucleotide sequences.

6. Conclusions

We trained Convolutional Neural Networks to predict Transcription Factor Binding Sites using a novel data representation, FGDR. The main difference between FGDR and the nucleotide sequence is that the former explicitly represents those chemical structures which are directly participating the DNA-protein interaction. We show that the performance of the binary classifier models are comparable to other relevant works and when the FGDR and Nuc. models are used in an ensemble setting, their results are better than the standalone predictors, furthermore they surpass other relevant methods in terms of AUC. An interesting prospect for future work is the combination of nucleotide and FGDR data. Several techniques could be applicable: concatenating the input matrices, training two convolutional blocks where one would receive nucleotide input, the other FGDR and merging the feature maps, or using the previous setup, the fully-connected layers could join the features learnt from the different input modalities. We are planning to explore these combinational approaches in the future.

7. Acknowledgements

Gergely Pap and László Tóth were supported by the National Research, Development and Innovation Office of Hungary through the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008). This study was supported by the Ministry of Innovation and Technology, grant number TUDFO/47138-1/2019-ITM. This work was a collaboration between the Bioinformatics Group of the Biological Research Centre, Szeged and the Institute of Informatics, University of Szeged. ZH and LT equally contributed to this study.

References

[1] B. Alipanahi, A. Delong, M. T. Weirauch és B. J. Frey, „Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nature Biotechnology,* %1. kötet33, p. 831–838, 8 2015.

[2] H. Zeng, M. D. Edwards, G. Liu és D. K. Gifford, „Convolutional neural network architectures for predicting DNA-protein binding,” *Bioinformatics,* %1. kötet32, p. i121–i127, 2/14.

[3] X. Pan és H.-B. Shen, „RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach,” *BMC Bioinformatics,* %1. kötet18, p. 136, 2017.

[4] H. R. Hassanzadeh és M. D. Wang, „DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins,” in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 2016.

[5] L. Fu, L. Zhang, E. Dollinger, Q. Peng, Q. Nie és X. Xie, „Predicting transcription factor binding in single cells through deep learning,” *bioRxiv,* p. 2020.01.14.905232, 1 2020.

[6] D. Quang és X. Xie, „FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data,” *Methods,* %1. kötet166, p. 40–47, 2019.

[7] W. Ma, L. Yang, R. Rohs és W. S. Noble, „DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding,” *Bioinformatics,* %1. kötet33, p. 3003–3010, 2/14.

[8] K. Adam, Z. Gyorgypal és Z. Hegedus, „DNA Readout Viewer (DRV): visualization of specificity determining patterns of protein-binding DNA segments,” *Bioinformatics,* %1. kötet36, p. 2286–2287, 12 2019.