PlotMI: visualization of pairwise interactions and positional preferences learned by a deep learning model from sequence data

Tuomo Hartonen[1], Teemu Kivioja[1], and Jussi Taipale[1,2,3,*]

[1]*Applied Tumor Genomics Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland*
[2]*Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom*
[3]*Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden*
[*]*Corresponding author: ajt208@cam.ac.uk*

March 13, 2021

Abstract

Deep learning models have in recent years gained success in various tasks related to understanding information coded in the DNA sequence. Rapidly developing genome-wide measurement technologies provide large quantities of data ideally suited for modeling using deep learning or other powerful machine learning approaches. Although offering state-of-the art predictive performance, the predictions made by deep learning models can be difficult to understand. In virtually all biological research, the understanding of how a predictive model works is as important as the raw predictive performance. Thus interpretation of deep learning models is an emerging hot topic especially in context of biological research. Here we describe plotMI, a mutual information based model interpretation strategy that can intuitively visualize positional preferences and pairwise interactions learned by any machine learning model trained on sequence data with a defined alphabet as input. PlotMI is freely available at `https://github.com/hartonen/plotMI`.

1 Introduction

Simultaneous rapid development of modern deep learning (DL) methods, computing hardware and high-throughput biological measurements has opened up avenues for succesful application of DL to different problems in the fields of biology and medicine [5]. DL models are especially well suited for analysis of data from large-scale DNA sequencing experiments as indicated for example by an early success story applying convolutional neural networks (CNNs) to learning binding motifs of transcription factors (TFs) [1].

Traditionally DL has been viewed as a "black box" method giving state-of-the-art performance at the cost of losing intepretability of the model. In many biological problems the ability to explain the decisions made by a machine learning (ML) model is at least as important as the accuracy of the predictions. Thus several methods for interpretation of DL models trained on biological sequence data have been developed.

In silico saturation mutagenesis (ISM) (used for example in [1]) is an intuitive approach that measures changes in model output produced by simulated mutations. In principle ISM can reveal contributions of any types of features but the computational cost of introducing mutations corresponding to more complex features limits the use of ISM to testing of specific hypotheses. Another widely employed strategy has been to weigh individual positions in selected input samples based on internal DL model metrics such as gradients or activation levels (e.g. [2, 11, 9]). Concentrating on individual

samples can however make it difficult to recognize more complicated features and thus approaches like sampling the maximum entropy distribution around sample inputs [7] and visualization of feature maps learned by the DL model [8] have been proposed.

A model-agnostic approach of interpreting DL model predictions is to feed random input to a pre-trained model and study features found from the high-scoring subset of the input (used e.g. in [3]). We propose a visualization approach that combines this idea with a measure of pairwise dependency between two variables, mutual information (MI) [10]. We use a pre-trained DL model to score sets of DNA (or other) sequences, filter out the part of the input based on model score and compute MI between pairwise position-specific k-mer distributions. The advantage of this approach is that it simultaneously samples a large space of high-scoring sequences and projects the pairwise interactions learned by the model into an easily interpretable two-dimensional visualization [13] that can highlight positions of important features as well as specific distances separating features learned by the model. Although this work demonstrates the method on DL models trained on DNA sequence, the implementation is agnostic to the type of the model or the sequence used to train the model.

2 Results

Workflow for MI analysis is shown in Figure 1 a), assuming a pre-trained DL model and the details of calculation of MI are described in Supplementary Materials. Main diagonal of the MI plot shows dependencies between k-mer distributions at adjacent positions, highlighting for example TF binding motifs in case of DNA sequences. Signal off the main diagonal corresponds to interactions between more distant positions. All analyses described in this work use 3-mers.

To illustrate the capability of plotMI to capture interactions learned by a DL model, we created a binary classification task where class 1 consist of DNA sequences sampled at random from uniform nucleotide background distribution and where binding motifs of TFs MAX and ELF5 were embedded into random positions in each of the sequences with exactly 50 bp between the start positions of the motifs. Class 0 sequences have similar background, but a single instance of either MAX-, or ELF5-motif was embedded into each sequence at random position (see Supplementary Materials). A convolutional neural network (CNN) was trained to classify between these two classes and subsequently used to score 10 million DNA sequences drawn from random uniform nucleotide background. All sequences obtaining a class 1 probability > 0.9 (29,442 sequences) were visualized with PlotMI (Figure 1 b)).

To demonstrate how MI plot can be used to discover specific pairwise interactions, we ran STREME [4] motif discovery on the high-scoring sequences and discovered a CACGTG-motif corresponding to the MAX-motif, one of the motifs in the embedded motif pair (see Supplementary Materials). We then created another set of 1 million random DNA sequences and embedded the CACGTG-motif in the middle of these sequences. The sequences were similarly scored using the CNN and sequences obtaining class 1 probability > 0.9 (13,331 sequences) were visualized with plotMI. Figure 1 c) shows that the CNN scores highly those sequences where another motif is approximately 50-60 bp off from the centered CACGTG-motif and that there is strong MI between these positions. A sequence logo of these 13,331 sequences revealed that this other motif corresponds to the ELF5-motif (Figure S1).

To study interactions learned from real datasets we downloaded 29,598 experimentally validated human transcription start site (TSS) positions from the Eukaryotic Promoter Database [6] and trained a CNN model to classify between ± 50 bp sequences around the TSS and shuffled versions of these sequences. We generated 1 million 100 bp long random DNA sequences with uniform nucleotide frequencies and scored them using the CNN yielding 49,261 sequences with promoter probability > 0.9. PlotMI visualization shows that the highest MI is found between adjacent positions right around the TSS, and that also a highly localized pairwise interaction between the TSS and the canonical TATA-box position at around 30 bp upstream from the TSS (Figure 1 d)) [6] is learned.

We further used plotMI to analyse interactions learned by a previously published model N-

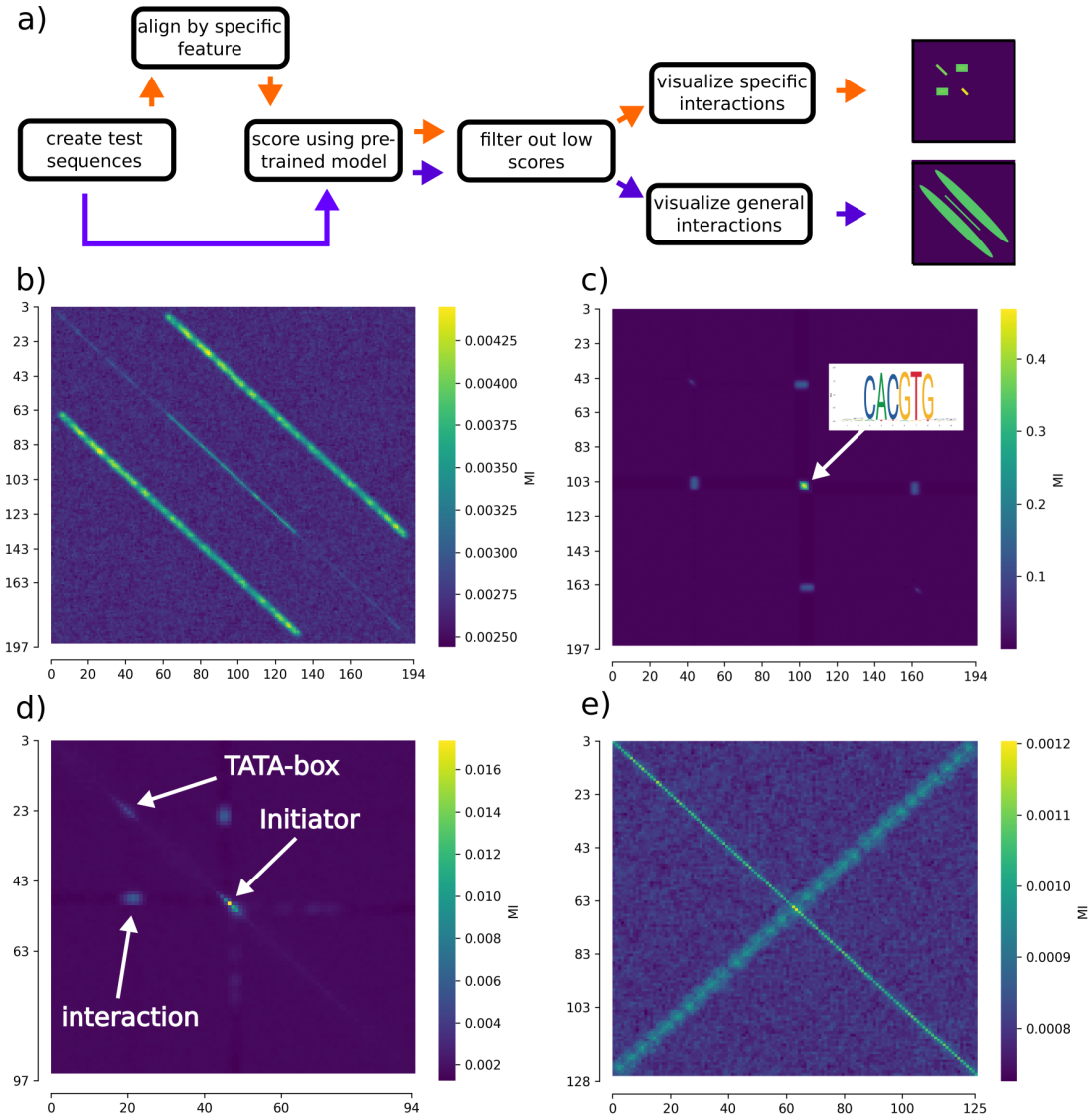

Figure 1: a) Two possible visualization workflows using plotMI. Mutual information (MI) can reveal the distance between interacting features if the visualized model is trained on unaligned data and the plotMI input sequences are also unaligned (blue arrows). If either the model is trained on aligned data, or the plotMI input sequences have been aligned based on a feature used by the model, MI can reveal the exact positions of the interacting features (orange arrows). The values at the x and y axes indicate positions along the input sequences. Panels b and c illustrate how alignment of input data based on a feature learned by a convolutional neural network (CNN) model alters MI plot. b) If the test data is not aligned plotMI will pick up interactions as lines parallel to the main diagonal of the MI matrix showing the preferred distance between the interacting features. MI plot shows that a CNN trained on a simulated dataset to classify between sequences with a specific MAX+ELF5 TF binding motif pair and sequences with either a MAX-, or ELF5-motif has learned the interaction embedded into the training data. c) If the test data is aligned based on a specific feature (or the model is trained on aligned data), interactions between features show off the main diagonal of the MI matrix. Motif discovery from the sequences visualized in panel b found a CACGTG-motif corresponding to the MAX-motif embedded into the training data of the CNN. Using the same CNN model to score sequences where this CACGTG-motif was embedded at position 100 shows that the CNN has learned a specific interaction with the CACGTG-motif and another motif that corresponds to the ELF5-motif embedded into the simulated training data (Figure S1). d) PlotMI reveals that a CNN classifier trained on 100bp long sequences centered at human genomic TSS positions has learned an interaction between the TSS and TATA-box region around 30 bp upstream from TSS. e) PlotMI visualization of N-score model trained to distinguish 131 bp long nucleosome binding DNA from non-nucleosomal DNA [12] shows that N-score has learned a periodic interaction pattern between the center of the nucleosome-favoring sequences and the flanking bases.

score, that is a wavelet-based logistic regression classifier developed to predict the nucleosome binding affinity of 131 bp long sequences in yeast [12]. We generated 500,000 sequences with uniform nucleotide background, scored them with N-score and selected the top 100,000 sequences for visualization. Figure 1 e) shows that, interestingly, N-score has learned that the middle position of the nucleosome binding sequences exhibits interactions with all other positions. From prior biological knowledge one would expect an approximately 10 bp periodic interaction manifesting as lines parallel to the main diagonal approximately 10 bp apart from each other. Main diagonal of the MI matrix shows that all positions share MI with their adjacent positions. Interestingly, plotMI visualization of the bottom 100,000 lowest-scoring sequences shows an even more prominent 10 bp periodic interaction between the 131 bp region middle point and the flanks (Figure S2).

3 Conclusions and Discussion

Here we apply the well-known measure of mutual dependency of random variables, mutual information (MI), to extract information about interactions learned by deep learning (DL) models. The aim of this approach is to project the information about pairwise and position-specific interactions learned by a DL model into a simpler model that can be intuitively visualized. This is done by feeding random input into a pre-trained model and using the model as a filter to select those input samples containing features learned by the model. These high-scoring sequences can be further mutated, or aligned based on some feature discovered using e.g. *de novo* motif mining to capture higher information content features. This type of additional cycles might be needed, as the visualization strategy described here can only highlight dependencies present in the input data. Some high information content features may effectively not be present in a completely random sample that is sufficiently small that it can be scored with a given ML model in reasonable time. Epistatic interactions can be captured by studying features present in low-scoring samples.

Importantly, the approach described here is not limited to MI visualization, but can be used to project information learned by a DL model to any other easily interpretable model. We show examples using binary classifies trained on DNA sequence data, but the workflow described here is applicable to regression models and other kinds of sequence data. Although the idea of interpreting a DL model by feeding it random input and studying features present in high-scoring input samples is not new, a general visualization tool like plotMI that can visualize pairwise interactions in an intuitive manner has to our knowledge not been previously described. The plotMI visualization script is freely available and can be used in conjunction with virtually any pre-trained ML model. We believe that the visualization approach described here will be a useful addition to the expanding toolbox of model interpretation methods available for scientists using ML to understand information encoded into DNA, protein and other types of biological sequence.

Acknowledgements

The authors would like to thank Drs. Esko Ukkonen, Fangjie Zhu and Kimmo Palin for helpful discussions regarding the mutual information visualization and Dr. Leena Salmela for help with selecting a nucleosome model for visualization. The authors wish to acknowledge CSC - IT Center for Science, Finland, for computational resources.

Funding

This work has been supported financially by the Academy of Finland grant to Center of Excellence in Tumor Genetics Research (grant no. 312042). *Conflict of interest:* None declared.

Author contributions

JT developed the original idea, TH implemented the plotMI program and performed the experiments. TH, TK and JT planned and wrote the manuscript.

References

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[3] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, pages 1–13, 2021.

[4] Timothy L Bailey. STREME: Accurate and versatile sequence motif discovery. *bioRxiv*, 2020.

[5] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H.S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. https://greenelab.github.io/deep-review/v/75f2dd8c61099a17235a4b8de0567b2364901e4d/.

[6] René Dreos, Giovanna Ambrosini, Romain Groux, Rouaïda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research*, 45(D1):D51–D55, 2017.

[7] Alex Finnegan and Jun S Song. Maximum entropy methods for extracting the learned features of deep neural networks. *PLoS computational biology*, 13(10):e1005836, 2017.

[8] Ge Liu, Haoyang Zeng, and David K Gifford. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC bioinformatics*, 20(1):1–14, 2019.

[9] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[10] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[11] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[12] Guo-Cheng Yuan and Jun S Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol*, 4(1):e13, 2008.

[13] Fangjie Zhu, Lucas Farnung, Eevi Kaasinen, Biswajyoti Sahu, Yimeng Yin, Bei Wei, Svetlana O Dodonova, Kazuhiro R Nitta, Ekaterina Morgunova, Minna Taipale, et al. The interaction landscape between transcription factors and the nucleosome. *Nature*, 562(7725):76–81, 2018.