NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Applied Mathematics and Informatics"

UDC 602

Research Project Report on the Topic:

The elaboration of a theoretical framework for the concurrent interpretation of various neural network approaches with the aim of identifying relevant features in the specific field of bioinformatics

Submitted by the Student:

group #БПМИ223, 3rd year of study Alaeva Ameliia Igorevna

Approved by the Project Supervisor:

Borevskiy Andrey Olegovich
Research Fellow
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Annotation

Today, the deep learning approach is widely used in many natural sciences and in bioinformatics in particular. In such complicated field as analysis of genomic data it is vital to understand on which factors model's decisions are based. Here stands the question of interpretability.

My team and I together aimed to develop a method that highlights the most relevant features, locates biological dependencies determined by neural network and could be applied to different problems.

We worked towards interpretation of neural networks in case of Z-DNA problem. The task of AI algorithm is to confirm or reject presence of Z-DNA in genome interval by taking into account omics data. We aimed to select the most important omics features on which the model relies while making a prediction. Together we developed a strong Convolutional and Graph deep learning models and tested Integrated Gradients, InputXGradients, Guided Backpropagation, Deconvolution interpretation methods on both architectures and Saliency, GNNExplainer on graph neural network. As a result, we identified the most important features which influence model's predictions in a high extent. They might have a biological significance regarding Z-DNA presence in genome. Moreover, we managed to select a larger group of features which is suitable for model's training with a negligible quality loss. This means, that less training data is required.

As a final result, we generalized our work to a universal interpretation framework which may be applied for omics data in genomic problems. Currently we are preparing a publication on this topic.

Keywords

Z-DNA is one of the many possible double helical structures of DNA. It is a left-handed double helical structure in which the helix winds to the left in a zigzag pattern, instead of to the right, like the more common B-DNA form.

Omics Data refers to data generated from high-throughput technologies used to study the various "omes" of an organism, such as the genome (all the genetic material), transcriptome (all the RNA molecules) and more.

Explainable AI (XAI) is a field of research within artificial intelligence (AI). The main focus is on the reasoning behind the decisions or predictions made by the AI algorithms, to make them more understandable and transparent.

Integrated Gradients, InputXGradients, Guided Backpropagation, Deconvolution, Saliency, GNNExplainer - XAI methods of model interpretation

Graph neural networks (GNN) are specialized artificial neural networks that are designed for tasks whose inputs are graphs

Convolutional neural network (CNN) is a regularized type of feed-forward neural network that learns features by itself via filter (or kernel) optimization.

Feature importance is the degree of influence of the particular feature on the predictions made by deep learning model.

1 Introduction

Deep Learning is a powerful tool for solving many issues in natural sciences, though the neural network is usually seen as a black box for the researchers. Meanwhile, it may be crucial or even more important in some cases to understand on which factors model's decisions are based. That is why, the field of study called Explainable AI (XAI) is gaining popularity. It focuses on the reasoning behind the predictions made by the AI algorithms, to make them more understandable and transparent. The main terms of XAI are Explainability and Interpretability.

There is a number of explainability and interpretability techniques. For instance, the process of reverse-engineering known as "mechanistic interpretability" or LIME (approximation locally a model's outputs with a simpler, interpretable model). We experimented with several interpretation methods and considered their benefits and drawbacks. Finally, we focused on gradient algorithms.

The goal of our research is to create a flexible framework suitable for omics data in genomic problems that locates biological dependencies determined by neural network.

We considered interpretation methods in case of the Z-DNA problem. Z-DNA is one of the many possible double helical structures of DNA. It is a left-handed structure in which the helix winds to the left in a zigzag pattern, instead of to the right, like the more common B-DNA form. It was first discovered in 1979 and still remains a mystery for scientists. However some studies show that people suffering from cancer or an Alzheimer disease have excessive amount of Z-DNA in genome. Based on the deep learning approach of the research [15] we designed our own high-quality models that identify presence of Z-DNA in genome interval taking omics features as an input. We interpretated both of them and extracted the most important features which may have biological significance.

The results of our interpretation framework applied to Z-DNA problem confirmed the success of approach. Additionally, we proved that a larger group of features (yet smaller than the initial dataset) may be extracted for sufficient model training with negligible quality loss.

Currently, we are preparing a publication on this topic.

2 Literature Overview

The first step before starting an interpretation and building a strong deep learning model is to analyze the subject field.

The Explainable AI methods started to develop in the early 1980s. [11] In the book published in 1986 [5], researchers for one of the first times dealt with the issues of explanations and transparency. However, as the models become stronger and more complicated, the need for their interpretability gets even greater [14]. For instance, Large Language Models achieved significant results in genomics considering contextual regions of different sizes at the DNA sequence level [2] [1]. The integration of omics data into Deep Learning approach can improve a model's prediction power and help to discover important associations between functional genomic elements and omics features. For example, multi-omics approaches became particularly important in cancer research. [19]

Here we are working with Z-DNA, which was first found in 1979 [12]. However, the researchers are still discovering much information about it. [18] [3] Z-DNA impacts the processes connected with chromatin regulation mainly and it was linked to both cancer and Alzheimer's disease. There were many attempts to identify Z-DNA presence in genome interval. The famous algorithm for predicting Z-DNA is statistically based and is called Z-HUNT [17]. Later it was outperformed in the paper [15] by deep learning model DeepZ. However, there was still a need for new powerful neural network architectures which perform with better accuracy and F1-score.

In a related paper [16] the formation of Z-DNA in promoters is discussed conserved between human and mouse using an ablation analysis with gradient boosting and PFI method. We, for our part, aimed to analyze which omics features are linked with the Z-DNA presence closely by using neural network's predictions. Interpretation algorithms [13] are suitable for this purpose. At present, there are a variety of XAI techniques.

Here we concentrate on Integrated Gradients [20], InputXGradients, Guided Backpropagation [4] , Deconvolution [6] for both our models and Saliency [22], GNNExplainer [21] for the graph model. The reasoning behind the methods selection and their peculiarities are discussed in the XAI Methods Overview section.

Using techniques listed above, we managed to extract the most important features for Z-DNA issue and to develop the Omics Interpretation Framework which can be used for different tasks.

3 Interpretation Prerequisites

The goal of interpretation is to make the model's decision process transparent. In our case it means to highlight the input features which are valuable for the network's prediction. It is highly likely that they represent a field of interest for the scientists or play an important role in an analytical solution.

The neural networks which have high quality metrics and strong prediction power are needed for the meaningful interpretation results. Weak models base their decisions on side rather than important features.

We developed two models which indicate the presence of Z-DNA in genome interval. First one is CNN model ConvMZC (ROC-AUC = 0.979 and F1 = 0.88). Second one is the graph based architecture GraphMZC(ROC-AUC = 0.958 and F1 = 0.81). As this paper is devoted to XAI methods, we will not discuss them here. More details will be given in our publication.

4 XAI Methods Overview

As was already mentioned we experimented with several XAI interpretation techniques. We focused on methods that indicate the extent at which features are contributing to a model's output. We have excluded some of them from the further study (e.g. LRP, Guided GradCAM). We decided to concentrate on the gradient methods and other closely connected techniques due to their theoretical robustness, ability to handle nonlinear interactions, and robustness to noise. They may lack efficiency in computation but they can provide the accuracy and faithfulness required for scientific research in genomics.

We have two different architectures: convolutional and graph neural networks. Integrated Gradients, InputXGradients, Guided Backpropagation, Deconvolution were applied to both of them; Saliency and GNNExplainer only to graph model. In this section we observe the methodology of every algorithm.

Integrated Gradients involves the calculation of gradients at each stage of the interpolation path. Formally, it is the integral of gradients with respect to inputs along the path from a given baseline to input. The integral can be approximated using a Riemann Sum or Gauss Legendre quadrature rule.

These gradients indicate the responsiveness of the model's prediction to infinitesimal changes in each feature at that precise point. This integration averages the influence of the feature over the whole range of values encountered during the process of interpolation. This provides a more

comprehensive and robust attribution score when compared to the scores provided by single-point gradients. The integrated gradients thus represent the mean impact of each feature on the model's prediction.

Saliency technique is based on returning the gradient of the output with respect to the input. This operation could be explained as taking a first-order Taylor expansion of the model at the input. In this case the gradients are seen as the coefficients of each feature in the linear representation of the network.

The feature importance is defined as an absolute value of these coefficients. Again the ability of gradients to show the sensitivity of the output with respect to changes in the input is used. Althoug Saliency method is quite simple, it has become a base for more complicated approaches such as Integrated Gradients and InputXGradients, DeepLift, Layer-Relevance Propagation.

InputXGradients is a further improvement of the saliency approach. The gradients of the output are taken with the respect to input and are multiplied by the input feature values. This operation can be understood as linear model: the gradients are coefficients of each input. The product of the input with a coefficient refers to the total contribution of the feature to the linear model's output.

Guided Backpropagation and **Deconvolution** calculate the gradient of the output with respect to the input, but the backward propagation of ReLU functions is modified so that only non-negative gradients are backpropagated. so that only non-negative gradients are backpropagated. In Guided Backpropagation, the ReLU function is applied to the input gradients, and in Deconvolution, the ReLU function is applied to the output gradients and directly propagated back. Both approaches were originally designed for the interpretation of convolutional networks and still are mainly used for CNN architectures.

GNNExplainer matches graph masks (adjacency matrices) that highlight which subgraphs and subsets of node features influence the most on the predictions of the GNN. To identify them an optimization problem in the form of maximization of mutual information between the distribution of predictions and the distribution of possible subgraphs and subsets of node features is solved.

5 Interpretation Pipeline

We developed an approach, which includes the XAI algorithms described above, to highlight the features playing a key role in the prediction made by DL model. We are providing step by step general **Interpretation Pipeline** and specify it in the case of Z-DNA problem:

1 **Choose an input tensor from the interval which contains target**

We prepare the dataset and keep the intervals which include Z-DNA as we are interested in their interpretation results.

2 **Get a prediction from well-pretrained model**

Get the prediction made by model as an input for interpretation method. The interpreted model should be effective in terms of evaluation metrics crucial for your problem. It is recommended to make observations by running interpretation pipeline several times using different well-pretrained models to achieve more accurate results.

In our case we focused on AUC-ROC and a high F1 score. We are using ConvMZC (F1 score = 88%) and GraphMZC (F1 score = 80%).

3 **Focus on True Positive regions only**

Only True Positive regions should be taken into account as they contain specific features playing a key role in successful model prediction we are looking for.

For us it means the regions which were predicted to contain Z-DNA and they actually do.

4 **Performing interpretation using XAI methods**

Conduct different interpretation methods which are suitable for the problem. We recommend using the Captum Library for PyTorch. The result of single input after one iteration of an interpretation algorithm is a tensor shaped [number_of_features, 1] which should be stored.

Among the methods chosen earlier Integrated Gradients, InputXGradients, Deconvolution and Guided Backpropogationsuitable are suitable for both CNN and Graph architectures. Additionally, we attempted GNNExplainer and Saliency for the Graph model.

Basically, the process of conducting an interpretation is similar within all methods we selected. First of all, we get a single tensor shaped [1950, 1] from the batch and make a prediction to indicate the indices which refer to the True Positive region. Then we pass an unchanged input tensor through an algorithm of interpretation with target parameter = 1

(focus on TP). We keep the indices which we identified as TP earlier. The output of an algorithm is an array shaped [1950, 1] = [ACTG + number_of_features, 1].

For Integrated Gradients and Input X Gradient its values could be negative (worsen prediction), neutral (no effect) or positive (improve prediction). For Deconvolution and Guided Backpropogation neutral (no effect) or positive (improve prediction) values are possible.

5 **Get an average interpretation score for a single algorithm**

Repeat steps 1 - 4 for the whole dataset to get an interpretation for every TP region. Then average outputs by axis=1 to get the tensor shaped [number_of_features, 1] which contains the average importance score of each feature.

Overall, 45201 tensors were interpreted: 9041 from test and 36160 from train data. Having the outcomes of interpretation we create tensor shaped [1950, 1] with the averaged importance values. It is the interpretation result of a single method for a particular model.

6 **Gather results for several XAI methods**

Get averaged feature importance scores (interpretation tensors) as an outcomes from several XAI methods for a deeper analysis. The indices of an interpretation outcome tensor correspond to the indices of the features in the input tensor. The greater the interpretation values are, the more importance the particular feature has.

We repeat the cycle for all selected methods using our both models. Finally, we have interpretation results of Integrated Gradients, Input X Gradient, Deconvolution and Guided Backpropogation both for Graph and CNN models.

6 Feature Extraction Process

Having the interpretation results of the XAI algorithms, the next step is to extract the relevant features. There are several ways to do it.

In our case indices from 0 to 3 correspond to the DNA nitrogen bases ACTG (A-Adenosine, C-Cytosine, T-Thymine, G-Guanine), so that they need to be included as features into the model's input by all means. That is why further we take into account slices of the original interpretation tensors from the 4th index to the end.

1 **The naive approach**

The straightforward solution is to range outcome of the method from the greatest values to the smallest keeping connection with the initial indices. Then one can select the most important features which have bigger interpretation values.

The difficulties occur when we interpret several models by using a couple of explanation algorithms. The ranged lists are highly likely differ for each of them. The intersection of these sorted lists will contain the features which were important regarding all methods for all models. On the one hand, it definitely increases robustness. On the other hand, we cannot control the number of relevant features to select. In case of little or no intersection the further analysis is impossible. That is why we used a more dependable approach.

2 **Statistical based approach**

Here we come to the **Ranking Stage** of our framework. The idea of relevance estimating is based on the calculation of percentage deviation.

1 Get the average value of an interpretation tensor for each XAI method.

2 For each item in tensor, compute the percentage deviation of its interpretation score from the corresponding mean for each XAI algorithm. So, the interpretation values are transformed into percent deviation scores now.

3 Compute the mean percentage deviation of each feature across all XAI methods. As a result, each feature has it's own ranking value.

In case of usage of several models, each feature has as many averaged values as the number of models. For instance, in our problem each feature has two values: one from the CNN and the other one from the GNN.

4 Sort the list of ranking values in descending order to get the most important features in the first place.

In case of using more than one model, several independent rankings are the outcome. This enables a comparison of the relative importance of each feature across the different neural network architectures.

As soon as we get the final ranking order of features, we find a solution to the given problem. That is because we can select the very first omics for the closer analysis regarding

biological meaning. At the same time we can consider the amount of k-first items (where k is a hyperparameter) for the efficient model training without a quality loss.

Note that one should keep the connection with the initial feature indices to be able to reconstruct the ranking results properly.

7 Results

Finally, we reached the main goal of our research. The Interpretation Pipeline and the Ranking Stage together form an **Omics Interpretation Framework**. It can be applied to a number of different tasks, for instance, in genomics. *It aims to evaluate the importance of each feature and to extract the most relevant of them. It sheds a light on the biological side of the problem and enables deeper analysis of the decision-making process of a neural network.*

Furthermore, the results in context of the Z-DNA problem will be discussed.

We conducted the ranking stage right after the interpretation pipeline. As a result, our algorithm extracts two groups of features from the initial omics data: a specific set of omics with biological meaning and a set suitable for training a high-quality model.

The 20 of 1000 top-extracted omics features obtained by our technique match with the omics selected in a related paper [16], where formation of Z-DNA in promoters is discussed conserved between human and mouse using an ablation analysis with gradient boosting and PFI method.

The quality of models which were trained on top-k extracted omics features (where k correspond to the number of selected features) is presented below. 7.1 7.2 As we see, the tuned hyperparameter k enables the better model performance rather than training on the whole feature dataset, which is a very crucial and successful result.

k, № of top features	ROC-AUC	F1-score
1950	0.9789	0.88
704	0.9755	0.879
504	0.9778	0.88
304	0.9771	0.8815
104	0.9739	0.8668
54	0.9748	0.8682

Figure 7.1: Performance of retrained GraphMZC architecture on Kouzine-Wu dataset with top-k features and interval size of 100 nucleotides.

k, № of top features	ROC-AUC	F1-score
1950	0.958	0.81
704	0.9584	0.8159
504	0.9596	0.8191
304	0.9597	0.8226
104	0.9623	0.8227
54	0.9612	0.8195

Figure 7.2: Performance of retrained ConvMZC architecture on Kouzine-Wu dataset with top-k features and interval size of 100 nucleotides.

Additionally, we tested our interpretation framework on the graph model which performed the same prediction task by using k-mers. These are the substrings of length k contained within a biological sequence which are composed of nucleotides (i.e. A, T, G, and C). The 'GCGC' sequence has top-interpretation scores. It matches the Chargaff's rules (biological theory regarding ratio of purine and pyrimidine bases) which confirms meaningfulness of our pipeline.

8 Software Implemented Framework

As a software outcome of our project, *we provide an implementation of our framework with a user-friendly interface.* We maintain the repository on Github [7] where several folders are available.

1 **CNN model framework folder** [8]

In this folder we provide the full interpretation framework including data processing, initial model training, interpretation pipeline, ranking stage, feature extraction and model training with the important features only. The main code gets started from the

"Omics_Interpretation_Framework_CNN.ipynb" notebook. It has a user-friendly interface and is ready to be modified for the given problem. The hyperparameters (e.g. width of the genome interval) may be tuned for the different tasks specially. The core functions (e.g. interpretation algorithms, ranking process) are defined in the separate ".py" files to make the main code more comprehensive.

2 **GNN model framework folder** [9]

The same as the previous folder but specified for GNN model.

3 **Interpretation folder** [10]

It contains Python notebooks devoted solely to the interpretation algorithms. They can be used for a deeper understanding of an Interpretation Pipeline conduction. Each XAI method for each model has its own notebook. The interpretation scores obtained for the Z-DNA problem are also published there.

The preview of framework notebooks is possible with Google Colab via links for CNN and GNN models.

9 Conclusion

In this project we :

1 Developed the flexible **Omics Interpretation Framework** suitable for omics data in genomic problems that extracts the most relevant features and locates biological dependencies determined by neural network

2 Created the powerful CNN- and GNN-based models ConvMZC and GraphMZC for Z-DNA identification

3 Applied framework to Z-DNA problem and extracted the relevant omics features

However, there are still more questions to be solved in the explainability of neural networks related to bioinformatics, in particular. We will continue our research on the topic.

References

[1] Eric Nguyen; Michael Poli; Marjan Faizi; Armin Thomas; Michael Wornow; Callum Birch-Sykes; Stefano Massaroli; Aman Patel; Clayton Rabideau; Yoshua Bengio; et al. "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution." In: *Advances in neural information processing systems, 36* (2024).

[2] Hugo Dalla-Torre; Liam Gonzalez; Javier Mendoza-Revilla; Nicolas Lopez Carranza; Adam Henryk Grzywaczewski; Francesco Oteri; Christian Dallago; Evan Trop; Bernardo P de Almeida; Hassan Sirelkhatim; et al. "Nucleotide transformer: building and evaluating robust foundation models for human genomics". In: *Nature Methods, pages 1–11* (2024).

[3] Patrick; Zhao Keji Liu Rui; Liu Hong; Chen Xin; Martha Kirby; O. Brownc. "Regulation of CSF1 Promoter by the SWI/SNF-like BAF Complex". In: *Cell Journal 106* (2001).

[4] Jost Tobias Springenberg; Alexey Dosovitskiy. *Striving for Simplicity: The All Convolutional Net*. 2015. URL: https://arxiv.org/abs/1412.6806.

[5] T. Winograd; F. Flores; F. F.Flores. "Understanding Computers and Cognition: A New Foundation for Design". In: *Intellect Books* (1986).

[6] Matthew D Zeiler; Rob Fergus. *Visualizing and Understanding Convolutional Networks*. 2013. URL: https://arxiv.org/abs/1311.2901.

[7] *Github: xAI_zdna_analysis*. URL: https://github.com/aameliig/xAI_zdna_analysis.

[8] *Github: xAI_zdna_analysis; CNN framework folder*. URL: https://github.com/aameliig/xAI_zdna_analysis/tree/main/cnn%20model%20framework.

[9] *Github: xAI_zdna_analysis; GNN framework folder*. URL: https://github.com/aameliig/xAI_zdna_analysis/tree/main/graph%20model%20framework.

[10] *Github: xAI_zdna_analysis; interpretation folder*. URL: https://github.com/aameliig/xAI_zdna_analysis/tree/main/interpretation.

[11] Mihály; Héder. ""Explainable AI: A Brief History of the Concept"". In: *ERCIM News (134): 9–10* (2023).

[12] Andrew Wang; Gary Quigley; Francis Kolpak. "Molecular structure of a left-Handed double helical DNA fragment at atomic resolution". In: *Nature 282* (1980).

[13] Captum: an open source library. *Algorithm Descriptions*. URL: https://captum.ai/docs/attribution_algorithms.

[14] Longo; Luca. *Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions*. 2024. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000794.

[15] Beknazarov Nazar; Jin Seungmin; Poptsova Maria. "Deep learning approach for predicting functional Z-DNA regions using omics data". In: *Scientific reportes 10* (2020).

[16] Nazar Beknazarov; Dmitry Konovalov; Alan Herbert Maria Poptsova. *Z-DNA formation in promoters conserved between human and mouse are associated with increased transcription reinitiation rates*. 2024. URL: https://www.nature.com/articles/s41598-024-68439-y#Sec19.

[17] G.J.; Rich A Ho P.S.; Ellison M.J.; Quigley. "A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences". In: *The EMBO Journal 5(10)* (1986).

[18] Christian Donners Marjo van der Vorst Emiel; Weber. "A Disintegrin and Metalloproteases (ADAMs) in Cardiovascular, Metabolic and Inflammatory Diseases: Aspects for Theranostic Approaches". In: *Thrombosis and Haemostasis* (2018).

[19] Lindong Jiang; Chao Xu; Yuntong Bai; Anqi Liu; Yun Gong; Yu-Ping Wang; and Hong-Wen Deng. "Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data." In: *NPJ precision oncology, 8(1):4* (2024).

[20] Mukund Sundararajan; Ankur Taly; Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. URL: https://arxiv.org/abs/1703.01365.

[21] Rex Ying; Dylan Bourgeois; Jiaxuan You. *GNNExplainer: Generating Explanations for Graph Neural Networks*. 2019. URL: https://arxiv.org/abs/1903.03894.

[22] Karen Simonyan; Andrea Vedaldi; Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. URL: https://arxiv.org/abs/1312.6034.