DNABERT, a linguistic approach for sequential predictions within Biology and Health

BACHELOR'S THESIS

DNABERT, a linguistic approach for sequential predictions within Biology and Health

BACHELOR'S THESIS

Student name/author
Daniël Kassab

Student nummer
S5795915

Location and date of completion
Buinen, 01-01-24

University
University of Groningen

Department
Faculty of Science and Engineering

Study
Bachelor Biology (Pre-master)

Academic supervisor
Dr. Anne de Jong

Summary

Performing predictions of elements as promoters, silencers, and enhancers in DNA- or RNA sequences within Biological- and Medical research is challenging due to their differences in functionality in different contexts. Although many current predictive models can predict with high accuracy, they cannot consider context. DNABERT addresses this limitation using a linguistic-based model in which DNA/RNA sequences are treated as a language.

This literature thesis discusses this new DNABERT model and addresses to what extent it has an impact on Biology and Health. Here, it was first identified whether DNABERT is revolutionary compared to current existing models. By comparing the obtained accuracy of the predictive models in previous studies with that of DNABERT, I concluded that DNABERT can obtain outstanding performance on splicing site prediction, and the highest accuracy but not outstanding can be obtained for promoter prediction. Consequently, I aimed to identify how DNABERT works, so that an understanding can be obtained that perhaps can be used for further optimization and customization. Therefore, the k-mer tokenization methods of DNABERT and the Byte-pair encoding were analysed. This was performed by adopting the described method for DNABERT from Ji et al. (2021) and DNABERT-2 from Zhou et al. (2023). From this analysis, it can be concluded that both methods are better than the existing methods for DNA/RNA predictions, but that BPE is the most promising. After that, there was a focus on promoter prediction using DNABERT (DNABERT-Prom) to obtain a clear insight into their process and how it was pre-trained. To obtain this information, the description of the DNABERT-Prom method from Ji et al. (2021) was adapted. Here, it could be identified that DNABERT-Prom was trained for predictions in *homo sapiens* using distal promoters that have a TATA-box present or absent. In addition, the EPDnew database was used to obtain the data for the promoters. After analysing DNABERT-Prom, I concluded that it is a highly efficient model to predict promoters within *homo sapiens*. Finally, I opted to provide a broader perspective of DNABERT to investigate how it can be applied within the field of Biology and Health. To analyse this, the described properties of DNABERT by Ji et al. (2021) were adapted and compared to the current constraints within Biology and Health. Here, I conclude that DNABERT is the most promising model for transcription regulation prediction in Biology and Health since it can address the required information of context, Furthermore. I conclude that DNABERT should also be the “first-choice” method to perform other types of DNA/RNA predictions, although their usage should never be a replacement for performing decisions within research and diagnostics. Although DNABERT is already a very sufficient model for prediction, further optimization and customization are still required to enlarge their contribution to sequential predictions within Biology and Health.

Table of contents

List of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte-Pair Encoding
CADD	Combined Annotation-Dependent Depletion
CNN	Convolutional neural network
DNN	Deep neural network
DSSP	Dictionary of Protein Secondary Structures
E.coli	*Escherichia coli*
EPDnew	Eukaryotic promoter database
GPT	Generative pre-trained transformer
IBPP	Image-based promoter prediction
LSTM	Long-Short Term Memory
MCC	Matthews correlation coefficient
ncRNA	non-coding RNA
ORF	Open reading frame
SNV	Single nucleotide variance

1. Introduction

1.1 Justification

A genome contributes to a tremendous amount of processes, ranging from transcribing essential proteins to keeping a household to an organism. The origin of these processes derives from sequences within the genome and are well organised in the form of promoters, enhancers, genes, and many more. At the same time, the genome is also very vulnerable to mutations, in which just a small error could already lead to disease. Interestingly, these organisational elements within sequences are highly similar between species, but the same element could function differently depending on the organism and the type of cell [44]. This demonstrates the essence of not only appreciating the order of nucleotides within a sequence itself, but it also has to be considered in their context [2]. This makes the performance of predictions of these elements a difficult task and many current predictive models (appendix, Table I) are imperfect due to their inability to analyse a sequence while taking context into account. DNABERT aims to overcome this limitation by using a linguistic-based approach as it treats these sequences as a language [2].

A predictive model that considers context is crucial for the optimal performance of predictions [2], [41], [32]. When successful, it can lead to major advancements in the understanding of many crucial biological processes, such as the transcription of genes [41]. In addition, it will also be crucial for Health since a better understanding of these biological processes can lead to better treatments. Because of this importance, this literature search aimed to unravel this new DNABERT in which an analysis was performed on whether this new model can be revolutionary for predictive tasks within Biology and Health. With this information, conclusions can be made on whether this new model should be further explored or another approach should be considered. Furthermore, it is also important to understand how context is extracted from a DNA/RNA sequence by DNABERT because this insight will provide opportunities for customization and further improvement of this model. Therefore, it was investigated how the k-mer tokenization method of DNABERT [2] and byte-pair encoding from DNABERT-2 [1] works. Finally, it is crucial to appreciate to what extent this new model can be useful for applications within Biology and Health because an impractical model would not be worth investigating. Hence, this study addressed this by investigating the current potentials of DNABERT and comparing these to the current limitations within Biology and Health [41], [42].

1.2 Problem statement

To what extent does DNABERT have an impact on Biology and Health; how does it work and how can it be used in research?

2. Research findings

2.1 DNABERT uses a linguistic approach to consider context while performing predictions

Predicting *cis*-regulatory elements within a genome is a great challenge in Biology. For example, predicting promoters, enhancers, and silencers for splicing is currently performed by various deep-learning approaches. However, a deep learning method called Bidirectional Encoder Representations from Transformers or BERT is the most reliable for finding *cis*-regulatory elements. This BERT is a deep learning program that focuses on language learning [2], [3]. A well-known example of a program that uses BERT is OpenAI GPT [3]. This language approach can also be used to read out DNA and this is exactly what DNABERT does. Hereby, it treats DNA/RNA sequences as a language. This approach can help to deal with the problem of polysemy. In human languages, this means that one word can have different meanings, for instance, the word “fly” can be labelled as a noun in one context (insect), but also as a verb in another (I fly by plane). This phenomenon can also be observed within DNA sequences [2], [4]. For example, the function of TATAAT is dependent on their location. For instance, it can be labelled as a starting point if it is located in a regulatory area, but it can also serve as a codon if it is present within genes for structure [4]. This shows the importance of including context to DNA/RNA sequences, and BERT is known to also bear this in mind while learning [2].

2.2 DNABERT is outstanding in splice site prediction and sufficient in promoter prediction compared to other models

The new DNABERT model has gained great attention recently since it indicates better performance compared to already existing models [2]. This raises the question of whether this new model is just one of many, or if it can obtain outstanding performance.

For many sequence predictions, there are a tremendous amount of models already available. For instance, the prediction of promoters can be performed by DeePromoter [5], CNNProm [6] and 70ProPred [7], [8], and the identification of transcription factor binding sites can be done using Deepsite [9], DESSO [10], or DanQ [2], [11]. To identify splice sites, various models are also available such as SPIDEX [12], [13], SpliceAI [12], [14], and SpliceFinder [15]. However, all models differ in their accuracy. To illustrate, CNNProm can obtain an F1-score of 0.948, while 70PromPred only achieved an F1-score of 0.897. This means that CNNProm can correctly identify 948 out of 1000 promoters, while for 70PromMed, this is only 897 (Appendix, Table I). The same variances in accuracy also apply to all described models for the detection of transcription factors and splicing sites. These differences in accuracy can originate from various factors. This means that the accuracy of a program is dependent on the year of development, method, and species. To indicate these differences, an overview of various programs is provided in Table I in the appendix.

DNABERT from 2021 and especially DNABERT-2 (2023) clearly shows to obtain one of the highest accuracy for most prediction tasks within sequences [1], [2]. This new model uses a different approach compared to the previous models in which it uses language modelling as the main focus. So can DNABERT achieve an F1-score above 0.900 for predictions of promoters, transcription factors, and splicing sites [2]. Still, the question remains whether DNABERT is revolutionary as a prediction model within Bioinformatics. To address this, the improvement of these prediction models across the years was analysed. Various previous studies (Appendix, table I) identified the extent of prediction accuracy across various programs. Since these models were released in different years, the progress of these models can be analysed and compared to those with DNABERT. As depicted in Figure 1, DNABERT indicates to follow the trend of improvement for promoter prediction, but is it outstanding for tasks such as splice site prediction. This concludes that DNABERT is revolutionising for some predictions, but not for others.

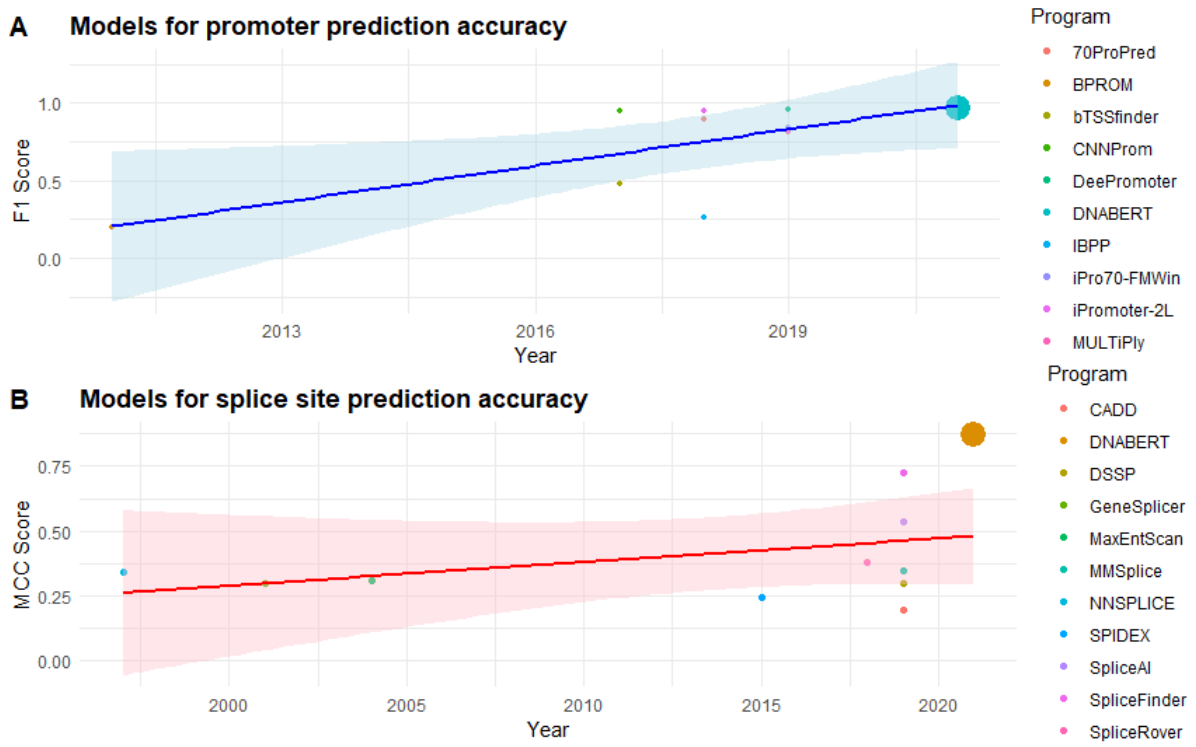

Figure 1. DNABERT has outstanding performance in splice site prediction and high performance in predicting promoters
Using the values for F1- and MCC accuracy, adapted from various studies (Appendix, table I), the trend of accuracy progression can be identified for promoter- and splice site prediction. In A, DNABERT (DNABERT-Prom, indicated as the largest blue dot) achieves an F1-score of 0.940, which is the highest among all other analysed promoter prediction programs. However, it follows the trend line (blue line) and falls within the confidence interval (light blue field), indicating that it is not outstanding. In B, DNABERT (DNABERT-Splice, largest red dot) also obtained the highest MCC-value of 0.871 [2], and it falls above and outside the confidence interval. This suggests that the predictiveness of DNABERT-Splice is revolutionary. Notice that the used values for these graphs are derived from different resources, meaning that it is uncertain whether these values were comparable in the first place. In addition, some programs lacked the required scoring values for this comparison and therefore, these were excluded from these graphs. Furthermore, different scoring methods were applied for splice site- and promoter prediction since only one kind of scoring method is provided for some programs. Finally, IBPP has a negative value of -0.03 [8], but this might be displayed as 0.03 due to inconveniences in processing negative values.

2.3 K-mer tokenization and especially BPE are very promising methods to obtain context from sequences

Previously, I discussed that DNABERT is a linguistic-based learning model. This means that DNA/RNA sequences will be treated as a language during the pre-training of this model. An important step in this pre-training is tokenization, which involves the segmentation of a piece of text. There are various methods present to do this but the main methods for BERT are Byte Pair Encoding (BPE) [33], and WordPiece [32], [34] which are both large language models [1]. However, the initial DNABERT uses k-mer tokenization to learn context from sequences [2].

In k-mer tokenization, a sequence is divided into a set of smaller fragments, also known as tokens. The size of these fragments depends on the size of the k-mer that is used. Hereby, the “K” indicates the number of nucleotides for each token. To illustrate this, consider the following sequence (CGGCTATAATCAG), which includes the previously described TATAAT. Using a k-mer size of 3 (3-mer), the sequence tokens will provide the following result: (CGG, GGC, GCT, CTA, TAT, ATA, TAA, AAT, ATC, TCA, CAG) [2]. To further encourage the k-mer tokenization for training based on the context of the entire sequence, one or a few of the tokens at random can be hidden, which is also known as masking. To demonstrate, “TAA” can be masked within our sequence by replacing it with “XXX” to get (CGG, GGC, GCT, CTA, TAT, ATA, XXX, AAT, ATC, TCA, CAG). However, it is important to understand that the way that tokens are identified is also problematic when it comes to truly hiding a sequence token. This is because the surrounding tokens (“ATA” and “AAT”) already reveal information about the hidden token, where the first two nucleotides for one token always match with the last two from the previous one. Since in our case, the XXX is surrounded by “ATA” and “AAT”, it already leaks that the hidden token must be “TAA”. This problem is referred to as the “leakage problem”. To combat this issue, non-overlapping tokenization can also be applied. In this case, the tokens will not contain overlapping (aligning) nucleotides [1]. Considering our example, the tokens in a 3-mer will now be identified as (CGG, CTA, TAA, TCA, G). Despite avoiding the issue of information leakage by the surrounding tokens, this approach is very sensitive for frameshifting when just one nucleotide is induced or removed. For instance, If a nucleotide “G” is induced in the beginning, it will result in completely different tokens (GCG, GCT, ATA, ATC, AG) [1], [35]. An overview of the two described methods of k-mer tokenization is depicted in Figure 2.

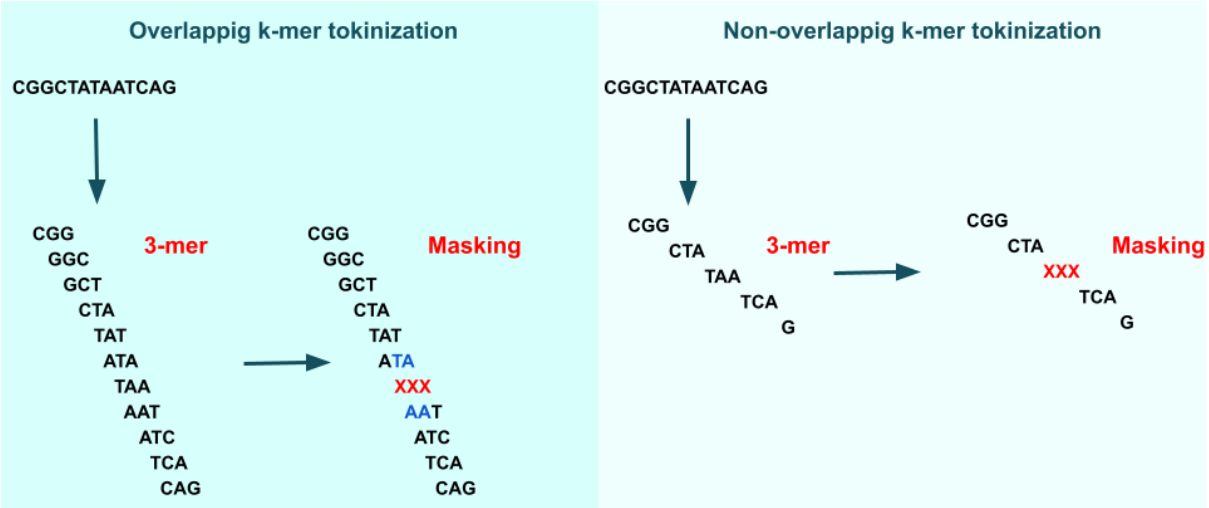

Figure 2. K-mer tokenization by DNABERT can be applied in an overlapping and non-overlapping setting A sequence can be divided into smaller fragments of size k. In the overlapping k-mer tokenization, it can be divided so that every next token has shifted one nucleotide in the downstream direction. In the example of 3-mer tokenization, every token consists of three nucleotides in which the first two overlap with the last two from the previous one. Consequently, one or few tokens can be hidden (masking) as depicted as "XXX", which replaces the "TAA" token. Notice that this masked token is leaked because the nucleotides from the surrounding tokens ("ATA" and "AAT", leaking nucleotides masked as dark blue) already indicate that the masked token must be "TAA". Another method is the usage of a non-overlapping setting where the nucleotides are divided into segments of size k (size 3 or 3-mer in this example) with no overlapping between the tokens. Notice that this time, the masked token is not leaked because the surrounding tokens do not contain any overlapping. However, this method is more vulnerable to frameshift since the insertion or deletion of just one nucleotide already results in a different set [1].

Since k-mer tokenization has been indicated to be suboptimal, a BPE approach for DNABERT tokenization was developed by Zhou et al. (2021) and it is named DNABERT-2 [1]. To understand the method of BPE, consider the previous sequence "CGGCTATAATCAG". This sequence can be used for BPE, and the entire pipeline is represented in Figure 3. The first step is to separate each nucleotide to get ("C", "G," "G", "C", "T", "A", "T", "A", "A", "T", "C", "A", "G"). This sequence of tokens will now be the starting vocabulary (initial corpus). After that, the algorithm will identify the occurrences of nucleotides that are paired together in the initial corpus. This can be done by observing the first token ("C") and then, combining it with the second one ("G") so that a token pair of "CG" can be made. This same principle can then be performed for the second token ("G") to the third ("C") to get a token pair of "GC". After performing this pairing to the entire sequence of tokens, the following can be obtained: ("CG", "GG", "GC", "CT", "TA", "AT", "TA", "AA", "AT", "TC", "CA", "AG"). Notice that some tokens now appear more than once in the sequence. After this, the number of appearances of each token pair in the sequence can now be listed ("CG":1, "GG":1, "GC":1, "CT"1, "TA":2, "AT": 2, "AA":1, "TC":1, "CA":1, "AG":1). From this list, the token pair with the highest number of appearances will now be added as a new token in the initial corpus. Now, the new corpus will be ("C", "G", "G", "C", "TA", "TA", "AT," "C", "A", "G"). Notice that "T" next to "A", and "A" next to "T" is now treated as a single token.
This process can then be replicated a few times until many tokens are identified within the corpus. For example, repeating this process with our new corpus library, the tokens ("CG", "GG", "GC", "CTA", "TATA", "TAAT", "ATC", "CA", "AG") can now be obtained. When listing the number of appearances again, it can be observed that each token has now only one appearance in the entire corpus. This means that our corpus cannot be tokenized further.

However, it is important to appreciate that only a small initial corpus was used in our example (CGGCTATAATCAG). In real applications of DNABERT-2, an initial corpus for pre-training can be used that contains over 100 million nucleotides [1], meaning that BPE tokenization can be performed many times. Considering our example, the current corpus (“CG”, “GG”, “GC”, “CTA”, “TATA”, “TAAT”, “ATC”, “CA”, “AG”) can now be used as a library to tokenize new sequences. For example, when the sequence “CGTATACACG” is tokenized again, the tokens (“CG”, “TATA”, ”CA, ”CG”) can now be obtained. This is also referred to as tokenizing a word [32], [33], [36].

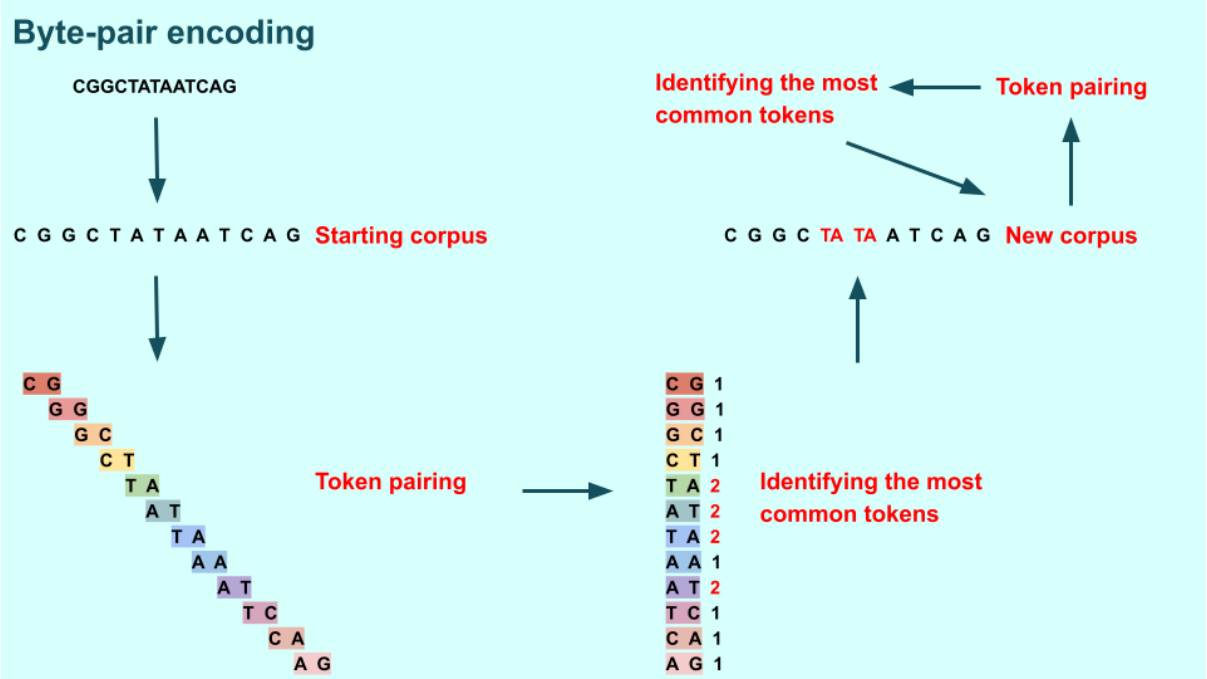

Figure 3. DNABERT-2 uses Byte-pair encoding (BPE) to obtain context from sequences
A sequence is first divided into all its single nucleotides and this will form the starting (initial) corpus. After that, each nucleotide will then be coupled to the next one and tokens will be created by moving one nucleotide downstream in the sequence, so that every first nucleotide of a pair will always have an overlap with the last one from the previous one This is depicted in “token pairing”. Consequently, the most frequent pair within all tokens will be identified, and these will be added to the initial corpus. Now, a new corpus is set in which the most common token will be treated as one single token. Finally, the process can then be repeated until many more tokens are identified [1].

2.4 DNABERT is the most promising for promoter prediction within their contexts.

Now that DNABERT is pre-trained to read genomic sequences as a language, it is capable of resolving many challenges within Biology and life sciences. For instance, it can deal with the issue of promoter prediction. Currently, there are three main strategies used to predict this. For instance, it can be identified based on the signal, where it exclusively focuses on the sequence components that correspond to the binding site of RNA polymerase.
Another approach is to look at CpG-islands because many promoters contain these regions within the genome. Finally, it can also be identified on a content basis, where k-mers are used and their frequency within a sequence is used to identify promoters [5]. However, all these strategies are not optimal for promoter prediction. A better model, namely DeePromoter [2], [5], was also introduced and has been shown to obtain better performance compared to the previous strategies. This model uses various deep learning approaches in their promoter identification, such as deep learning for recognizing patterns within images (Convolutional Neural Network, CNN) [37] combined with a model (Long short-term memory, LSTM) that can make predictions in text based on stored information from previous data [5], [38]. Despite that DeepProm is already very sufficient in predicting promoters, the new DNABERT is the most promising for promoter identification [2].

The strategy that DNABERT-Prom uses to predict genomes in *homo sapiens* is accurately described in Ji et al. (2021) [2]. First, the dataset can be trained on known proximal promoters in which some contain the TATA-box and some in which this box is absent.
To obtain this data, the EPDnew database [39] can be applied. For each promoter (TATA-box present or absent), a sequence of 10.000 nucleotides can be obtained in which the 5000 nucleotides will be upstream of the transcription start site and 5000 downstream.
After training, promoters can now be predicted by taking a TATA-box-containing sequence of 300 nucleotides of which 249 are downstream of the transcription start site and 50 upstream.

2.5 DNABERT can contribute to a better understanding of transcription regulation and the expression of oncogenes/tumour suppressor genes

As described previously, DNABERT has been shown to provide a higher accuracy in various predictions compared to other existing models. This is crucial for the improvement of research within Health and Biology. However, this is not what makes DNABERT revolutionary since an increased performance can also be obtained by switching to different deep learning approaches [40] or, based on own interpretation, by providing more training data. Although many predictive programs can accurately perform many common predictive tasks, some issues cannot be resolved by their prediction method. This is because sequences contain properties that are similar to human languages and therefore, these can only be addressed by linguistic models such as (DNA)BERT [2]. By all means, DNABERT contains unique properties to resolve certain predictive issues within Biology and Health that could not be addressed by other models [2].

A great example of the property of DNABERT to learn context is to better understand the regulation of transcription. This is because the regulation relies not only on the DNA sequence itself but also on the binding of transcription factors; histone modifications and the folding of chromatins. This means that regulation is dependent on their entire context [41]. Since DNABERT can achieve high performance in analysing sequences in their context, this new model can lead to a significant improvement in understanding how the elements of transcription collaborate [2].

Within the regulation of transcription, non-coding RNAs (ncRNA) also play an essential role, but they are poorly understood. Here, they have been shown to bind to the DNA during the regulation of transcription [41]. These ncRNAs are also associated with cancer in which various types of it contribute to the expression of tumour suppressor- and oncogenes [42]. To better understand the mechanism of these ncRNAs, DNABERT can provide a better understanding of how they work in their context.

3. Discussion and conclusion

3.1 Discussion

Performing predictions for promoters, transcription factors, and splicing sites within DNA/RNA sequences is crucial for biological- and medical research, but performing this is suboptimal because the function of these elements is dependent on their context [2]. Currently, many models from various studies are widely used to address these predictions (appendix, Table I), but these are unable to provide important information for context. Therefore, DNABERT is introduced to resolve this contextual issue by performing predictions based on language modelling [2]. For this reason, I opted to know how DNABERT with this new modelling approach works and to what extent it has an impact on Biology and Health. To investigate the potential of this new model, it was first compared to existing models for prediction with different approaches (appendix, Table I). From the results, it showed that DNABERT is outstanding in predictions for splicing sites, but not for promoter identification. This indicates that other bioinformatical predictions with DNABERT (enhancers, transcription factor-DNA, etc.) might also not obtain outstanding performance compared to the other models, but further analysis of these predictions is needed to determine this.
However, DNABERT should still be the model of choice since it can provide additional context which is crucial for predictions [2].

After the identification of the impact of DNABERT, I opted to understand how their model works to be able to use- and optimise it. Therefore, the mechanism of k-mer tokenization and BPE in DNABERT [1], [2] was explored and it can be concluded that both methods are useful for performing DNA/RNA sequential predictions, but that BPE would be the method of choice. This is because Toraman et al. (2023) [32] discussed that k-mer tokenization faces issues of vulnerability for SNVs and the leakage of information for hidden sequences. Having said that, the usage of linguistics as a basis for sequential predictions in general seems more suited for sequence predictions than a statistical approach due to the linguistic behaviour of DNA sequences [2]. However, it can be argued that a mathematical approach would be even better. To illustrate, Athens & José (2005) [43] were able to map coding from ncDNA by treating the sequence as a mathematical language. Here they considered all purines as “-1” and pyrimidines as “1” and when reading out the sequence from upstream to downstream, a sequence as “AAGC" could be read out as (“-1,-1,-1,1”). When plotting these values, a graphical image can be obtained that represents the sequence [43]. This suggests that such an approach could be used as a base for modelling and that perhaps curtain sequential elements in different contrasts could provide different graphical images on this method. Another option could be to apply a base-4 numbered system (real numbers of only 0,1,2 and 3) for “ATCG” where (A=0, T=1, C=2, and G=3), and by constructing corresponding formulas, it might be possible that sequential elements in their context might be explained mathematically.

After the exploration of the DNABERT method, I wanted to investigate how it can be applied to various tasks within Biology and Health. By analysing the properties of DNABERT and comparing it to common difficulties in research, it could be identified that it can be used to achieve a better understanding of the mechanism of transcription regulation [1], [41], and since this also plays a role in the development of tumours [42], DNABERT also contributes to cancer research. In this literature research, only a few applications of DNABERT were described, but it will also most likely be an optimal model to describe numerous other applications.

Finally, I strongly recommend using DNABERT only as an additional predictive program for research and diagnostics. Performing decisions should always be made based on the expertise of the researcher and predictive models such as DNABERT should not serve as a replacement.

3.2 Conclusion

DNABERT is a linguistic-based model and it should be the “first-choice” model for performing DNA/RNA sequential predictions due to its capability to perform predictions while considering context. More specifically, DNABERT-2 should be used since it achieves even better performance than DNABERT. Despite the high accuracy of other programs, DNABERT's unique property of predicting with context will make the predictions more robust, and it will provide possibilities for Medical and Biological research in the understanding of various biological and medical processes, which is currently limited by the absence of context within DNA/RNA prediction. However, I also conclude that it should never be used as a replacement to perform decisions within research and diagnostics. Finally, future studies should investigate the capabilities of DNABERT for more applications, and it should be further optimised and customised for increased performance.

4. Literature list

[1] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, en H. Liu, 'DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome', jun. 2023, accessed at: 14 november 2023. [Online]. Available at: http://arxiv.org/abs/2306.15006

[2] Y. Ji, Z. Zhou, H. Liu, en R. V. Davuluri, 'DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome', *Bioinformatics*, vol. 37, nr. 15, pp. 2112-2120, aug. 2021, doi: 10.1093/bioinformatics/btab083.

[3] J. Devlin, M.-W. Chang, K. Lee, en K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. arXiv, 24 mei 2019. Accessed: 15 november 2023. [Online]. Available at: http://arxiv.org/abs/1810.04805

[4] B. Enguix, 'Natural language and the genetic code: from the semiotic analogy to biolinguistics'.

[5] M. Oubounyt, Z. Louadi, H. Tayara, en K. T. Chong, 'DeePromoter: Robust Promoter Predictor Using Deep Learning', *Front. Genet.*, vol. 10, p. 286, apr. 2019, doi: 10.3389/fgene.2019.00286.

[6] R. K. Umarov en V. V. Solovyev, 'Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks.', *PloS One*, vol. 12, nr. 2, p. e0171410, 2017, doi: 10.1371/journal.pone.0171410.

[7] W. He, C. Jia, Y. Duan, en Q. Zou, '70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features', *BMC Syst. Biol.*, vol. 12, nr. Suppl 4, p. 44, apr. 2018, doi: 10.1186/s12918-018-0570-1.

[8] M. H. A. Cassiano en R. Silva-Rocha, 'Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations', *mSystems*, vol. 5, nr. 4, pp. e00439-20, aug. 2020, doi: 10.1128/mSystems.00439-20.

[9] Y. Zhang, S. Qiao, S. Ji, en Y. Li, 'DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding', *Int. J. Mach. Learn. Cybern.*, vol. 11, nr. 4, pp. 841-851, apr. 2020, doi: 10.1007/s13042-019-00990-x.

[10] A. M. Khamis *e.a.*, 'A novel method for improved accuracy of transcription factor binding site prediction', *Nucleic Acids Res.*, vol. 46, nr. 12, p. e72, jul. 2018, doi: 10.1093/nar/gky237.

[11] D. Quang en X. Xie, 'DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences', *Nucleic Acids Res.*, vol. 44, nr. 11, p. e107, jun. 2016, doi: 10.1093/nar/gkw226.

[12] T. V. Riepe, M. Khan, S. Roosing, F. P. M. Cremers, en P. A. C. 't Hoen, 'Benchmarking deep learning splice prediction tools using functional splice assays', *Hum. Mutat.*, vol. 42, nr. 7, pp. 799-810, jul. 2021, doi: 10.1002/humu.24212.

[13] H. Y. Xiong *e.a.*, 'RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease', *Science*, vol. 347, nr. 6218, p. 1254806, jan. 2015, doi: 10.1126/science.1254806.

[14] K. Jaganathan *e.a.*, 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, vol. 176, nr. 3, pp. 535-548.e24, jan. 2019, doi: 10.1016/j.cell.2018.12.015.

[15] R. Wang, Z. Wang, J. Wang, en S. Li, 'SpliceFinder: ab initio prediction of splice sites using convolutional neural network', *BMC Bioinformatics*, vol. 20, nr. 23, p. 652, dec. 2019, doi: 10.1186/s12859-019-3306-3.

[16] B. Alipanahi, A. Delong, M. T. Weirauch, en B. J. Frey, 'Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning', *Nat. Biotechnol.*, vol. 33, nr. 8, pp. 831-838, aug. 2015, doi: 10.1038/nbt.3300.

[17] J. Zhou en O. G. Troyanskaya, 'Predicting effects of noncoding variants with deep learning-based sequence model', *Nat. Methods*, vol. 12, nr. 10, pp. 931-934, okt. 2015, doi: 10.1038/nmeth.3547.

[18] D. R. Kelley, J. Snoek, en J. L. Rinn, 'Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks', *Genome Res.*, vol. 26, nr. 7, pp. 990-999, jul. 2016, doi: 10.1101/gr.200535.115.
[19] 'Salamov: Metagenomics and its applications in agriculture... - Google Scholar'. Accessed: 28 november 2023. [Online]. Available at: https://scholar.google.com/scholar_lookup?title=Metagenomics+and+its+applications+in+agriculture,+biomedicine+and+environmental+studies&author=V+Solovyev&author=A+Salamov&publication_year=2011&
[20] I. A. Shahmuradov, R. Mohamad Razali, S. Bougouffa, A. Radovanovic, en V. B. Bajic, 'bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli', *Bioinforma. Oxf. Engl.*, vol. 33, nr. 3, pp. 334-340, feb. 2017, doi: 10.1093/bioinformatics/btw629.
[21] S. Wang, X. Cheng, Y. Li, M. Wu, en Y. Zhao, 'Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns', *Sci. Rep.*, vol. 8, nr. 1, p. 17695, dec. 2018, doi: 10.1038/s41598-018-36308-0.
[22] B. Liu, F. Yang, D.-S. Huang, en K.-C. Chou, 'iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC', *Bioinforma. Oxf. Engl.*, vol. 34, nr. 1, pp. 33-40, jan. 2018, doi: 10.1093/bioinformatics/btx579.
[23] M. Zhang *e.a.*, 'MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters', *Bioinforma. Oxf. Engl.*, vol. 35, nr. 17, pp. 2957-2965, sep. 2019, doi: 10.1093/bioinformatics/btz016.
[24] M. S. Rahman, U. Aktar, M. R. Jani, en S. Shatabda, 'iPro70-FMWin: identifying Sigma70 promoters using multiple windowing and minimal features', *Mol. Genet. Genomics MGG*, vol. 294, nr. 1, pp. 69-84, feb. 2019, doi: 10.1007/s00438-018-1487-5.
[25] M. Pertea, X. Lin, en S. L. Salzberg, 'GeneSplicer: a new computational method for splice site prediction', *Nucleic Acids Res.*, vol. 29, nr. 5, pp. 1185-1190, mrt. 2001, doi: 10.1093/nar/29.5.1185.
[26] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, en M. Kircher, 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Res.*, vol. 47, nr. D1, pp. D886-D894, jan. 2019, doi: 10.1093/nar/gky1016.
[27] T. Naito, 'Predicting the impact of single nucleotide variants on splicing via sequence-based deep neural networks and genomic features', *Hum. Mutat.*, vol. 40, nr. 9, pp. 1261-1269, sep. 2019, doi: 10.1002/humu.23794.
[28] G. Yeo en C. B. Burge, 'Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals', *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, vol. 11, nr. 2-3, pp. 377-394, 2004, doi: 10.1089/1066527041410418.
[29] J. Cheng *e.a.*, 'MMSplice: modular modeling improves the predictions of genetic variant effects on splicing', *Genome Biol.*, vol. 20, nr. 1, p. 48, mrt. 2019, doi: 10.1186/s13059-019-1653-z.
[30] M. G. Reese, F. H. Eeckman, D. Kulp, en D. Haussler, 'Improved splice site detection in Genie', *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, vol. 4, nr. 3, pp. 311-323, 1997, doi: 10.1089/cmb.1997.4.311.
[31] J. Zuallaert, F. Godin, M. Kim, A. Soete, Y. Saeys, en W. De Neve, 'SpliceRover: interpretable convolutional neural networks for improved splice site prediction', *Bioinforma. Oxf. Engl.*, vol. 34, nr. 24, pp. 4180-4188, dec. 2018, doi: 10.1093/bioinformatics/bty497.
[32] C. Toraman, E. H. Yilmaz, F. Şahinuç, en O. Ozcelik, 'Impact of Tokenization on Language Models: An Analysis for Turkish', *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, nr. 4, pp. 1-21, apr. 2023, doi: 10.1145/3578707.
[33] R. Sennrich, B. Haddow, en A. Birch, 'Neural Machine Translation of Rare Words with Subword Units', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk en N. A. Smith, Red., Berlin, Germany: Association for Computational Linguistics, aug. 2016, pp. 1715-1725. doi: 10.18653/v1/P16-1162.
[34] M. Schuster en K. Nakajima, 'Japanese and Korean voice search', in *2012 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), mrt. 2012, pp. 5149-5152. doi: 10.1109/ICASSP.2012.6289079.
[35] C. Liang *e.a.*, 'Rethinking the BERT-like Pretraining for DNA Sequences'. arXiv, 11 oktober 2023. Accessed: 18 november 2023. [Online]. Available at: http://arxiv.org/abs/2310.07644
[36] 'Byte-Pair Encoding tokenization - Hugging Face NLP Course'. Accessed: 19 november 2023. [Online]. Available at: https://huggingface.co/learn/nlp-course/chapter6/5
[37] K. O'Shea en R. Nash, 'An Introduction to Convolutional Neural Networks'. arXiv, 2 december 2015. Accessed: 20 november 2023. [Online]. Available at: http://arxiv.org/abs/1511.08458
[38] Z. Huang, W. Xu, en K. Yu, 'Bidirectional LSTM-CRF Models for Sequence Tagging'. arXiv, 9 augustus 2015. Accessed: 20 november 2023. [Online]. Available at: http://arxiv.org/abs/1508.01991
[39] R. Dreos, G. Ambrosini, R. Cavin Périer, en P. Bucher, 'EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era', *Nucleic Acids Res.*, vol. 41, nr. Database issue, pp. D157-164, jan. 2013, doi: 10.1093/nar/gks1233.
[40] J. Ahmad, H. Farman, en Z. Jan, 'Deep Learning Methods and Applications', in *Deep Learning: Convergence to Big Data Analytics*, M. Khan, B. Jan, en H. Farman, Red., in SpringerBriefs in Computer Science. , Singapore: Springer, 2019, pp. 31-42. doi: 10.1007/978-981-13-3459-7_3.
[41] J. H. Gibcus en J. Dekker, 'The context of gene expression regulation', *F1000 Biol. Rep.*, vol. 4, p. 8, apr. 2012, doi: 10.3410/B4-8.
[42] S. Zhang *e.a.*, 'Oncogenic MORC2 in cancer development and beyond', *Genes Dis.*, vol. 11, nr. 2, pp. 861-873, jul. 2023, doi: 10.1016/j.gendis.2023.05.010.
[43] J. Athens en M. José, 'Mathematical properties of DNA sequences from coding and noncoding regions', *Rev. Mex. Física*, vol. 51, apr. 2005.
[44] M. Nirenberg e.a., 'RNA codewords and protein synthesis, VII. On the general nature of the RNA code.', Proc. Natl. Acad. Sci., vol. 53, nr. 5, pp. 1161-1168, mei 1965, doi: 10.1073/pnas.53.5.1161.

5. Foreword and afterword

This literature research was conducted under the supervision of Dr. Anne de Jong from the University of Groningen. Therefore, I would like to thank Dr. Anne de Jong for his support. The topic of DNABERT was considered due to the high demand for an overview of this new model that also would be understandable for biologists with little to no bioinformatical knowledge. I would further thank the University of Groningen in general for facilitating the opportunity to perform this literature study.

6. Supplementary

6.1 Datasets Figure 1

Link to datasets package for figure 1 (ModelsComparison)

6.2 Additional information of the analysed prediction programs

Table I. Overview of the currently used programs for promoter- and splicing site prediction

Program	Prediction	Description	Reference
70ProPred	Promoter	Predicting sigma 70 promoters in prokaryotes	[7], [8]
BPROM	Promoter	Predict promoters upstream from predicted ORFs in microbes	[8], [19]
btSSfinder	Promoter	Predicting various sigma factor promoters in cyanobacteria of *Escherichia coli* (*E.coli*)	[8], [20]
CNNProm	Promoter	Convolutional neural network (CNN) based promoter prediction for prokaryotes and eukaryotes	[6], [8]
DeePromoter	Promoter	Promoter prediction for eukaryotes (high accuracy for human and mouse), based on a combination of CNN and Long-Short Term Memory (LSTM)	[2], [5]
DNABERT (DNABERT-Prom)	Promoter	Promoter prediction for human sequences, based on the Bidirectional Encoder Representations from Transformers (BERT) language modelling	[2]
DNABERT-2	Promoter	Improved version of DNABERT-Prom, that not only predicts promoters in human sequences, but also in various other prokaryotes and eukaryotes	[1]
IBPP	Promoter	Predicts promoters based on comparing sequences to an "image" (a developed sequence of a promoter). Trained on the sigma 70 promoter of *E.coli*	[8], [21]
iPro70-FMWin	Promoter	Prediction of promoters based on the E.coli sigma 70 promoter	[8], [24]
iPromoter-2L	Promoter	To predict promoters in *E.coli* in a system of two layers. The first layer determines whether a sequence is a promoter and the second one classifies to what sigma promoter it belongs to.	[8], [22]
MultiPly	Promoter	Predicting promoters of all sigma factors of *E.coli*	[8], [23]
CADD	Splice site	Predicting splicing sites using a standard machine learning method. It was trained on many SNVs.	[12], [26]
DNABERT (DNABERT-Splice)	Splice site	Language modelling based (BERT) prediction of splicing sites in humans	[2]

DNABERT-2	Splice site	Improved version of DNABERT-Splice that allows for splicing site prediction in other eukaryotes and prokaryotes	[1]
DSSP	Splice site	Splicing site prediction based on a combination of CNN and LSTM	[12], [27]
GeneSplicer	Splice site	Predicting splicing sites in eukaryotes based on Markov modelling and a decision tree. Furthermore, it was trained on various genes within plants and humans	[12], [25]
MaxEntScan	Splice site	Predicting RNA splicing based on maximum entropy. For the prediction, splicing sites of humans were compared to decoys	[12], [28]
MMSplice	Splice site	A neural network-based approach to predict variances in splicing sites. The training was performed on all the nucleotides that are present within the structure of exons and introns	[12], [29]
NNSPLICE	Splice site	Predicting splicing sites using neural networks and a hidden Markov model. The training was performed on the exons of humans, which are publicly available on GenBank	[12], [30]
SPIDEX	Splice site	A classical machine learning method to predict splicing sites. The method that SPIDEX uses is "Bayesian modelling", which is a statistical-based model.	[12], [13]
SpliceAI	Splice site	A deep learning approach for predicting non-coding variants of pre-mRNA	[12], [14]
SpliceFinder	Splice site	Predicting splicing sites using CNN, and training was performed on the human genome	[2], [15]
SpliceRover	Splice site	Splicing site prediction using CNN, and it was trained on humans and plants	[31]