Special Proceeding: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 177-184

Opportunities and Challenges in Statistics and Data Science

Yanrong Ji[1] and Ramana V. Davuluri[2]
[1]*Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.*
[2]*Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA.*

Received: 16 August 2021; Revised: 03 September 2021; Accepted: 06 September 2021

Abstract

Statisticians play critical roles in various fields of study that deal with data and decision making in the face of uncertainty. The rise of recent 'Data Science' offer exceptional opportunity to Statisticians because of the universal relevance of statistical methods in the interpretation of data. However, there are some challenges to become an impactful "Data Scientist" and/or "Statistician". The core of this article will be a summary of some recent research projects, through which we wish to demonstrate that statistics together with information technology makes an essential contribution to the emerging fields of 'precision medicine' and 'genomics' research. Finally, we offer our thoughts on how deep-learning methods might best be adapted for data-scarce scenarios to achieve exceptional performance.

Key words: Bioinformatics; Genomics; Machine Learning; Deep Learning.

1. Introduction

The genetic alphabet – consisting of just four letters A, C, G, and T; forms a simple and elegant language by combining redundancy and utility in the genomes. For example, a gene is a sentence of sequence of words (for example, exons and introns, like "GCTGCTGCAGAA…CTGCCTAGA", which has a certain biological meaning or function. In the most straightforward case, a gene is translated into a specific protein, made of amino acids. The information in a gene is organized into three-letter words called codons, 64 possible triplets from the 4 genetic letters, which translate into 20 different amino acids and 1 stop codon, allowing redundancy. Deciphering the language of DNA for these and other hidden instructions in translating the information from DNA to RNA and RNA to protein has been one of the ultimate goals of biological research (Portin, 2014). Statistical machine learning methods and computer systems have played critical role in interpreting the language of DNA in the human genome.

2. Deciphering the Language of Protein-coding DNA

In early days of the human genome sequencing phase, most of the gene prediction programs ((Zhang, 2002; Davuluri and Zhang, 2003)) have been mainly trained to predict coding exons, stretches of gene sequences that get translated into sequence of amino acids forming a specific protein. The crux of these programs are statistical models that calculates the

Corresponding Author: Ramana V. Davuluri;
Email: Ramana.Davuluri@stonybrookmedicine.edu

probability of a given stretch of DNA is coding exon or not. For example, GenScan (Burge and Karlin, 1997) and Genie (Kulp *et. al.*, 1996) use hidden Markov models, Grail (Xu *et. al.*, 1994) uses Neural Networks and MZEF applied quadratic discriminant functions. While the gene predictions programs were successful in accurately prediction protein-coding exonic regions, prediction of gene-promoters and non-coding first exons, remained as a highly complex problem and critical gap in gene-prediction for several years during the human genome sequencing phase. FirstEF program was developed to identify first exons and promoters by scanning human DNA sequences. It operates by finding every potential first splice-donor site (using donor QDF – *quadratic discriminant function*) and promoter (using promoter QDF) and then calculating the probability that the intervening sequence is a first exon (using exon QDF). The power of FirstEF lies in its ability to identify first exons that are either CpG related or non-CpG-related, using different sets of classification models. The details of FirstEF algorithm are described in Davuluri *et al.* (Davuluri *et. al.*, 2001) and its application on human genome sequences is explained in (Davuluri, 2003). These creative and groundbreaking approaches facilitated the prediction and annotation of Pol-II promoters in human and mouse genomes.

3. Deciphering the Language of Regulatory DNA

While the genetic code that explains how DNA is translated into proteins is universal, the gene regulatory code that determines how and when the genes are expressed varies across different cell-types and tissues (Nirenberg *et. al.*, 1965). Over the past decade, Next-Generation sequencing (NGS) technologies have accelerated our ability to generate numerous epigenomic and transcriptomic datasets (Dunham *et. al.*, 2012). These datasets collectively facilitated genome-wide identification of gene regulatory regions in an unprecedented way and unveiled the complexity of the human genes and their regulation. The exponential growth of the multi-omics data, computational analysis of large datasets has become commonplace in the study of human biology and disease.

For example, in one of our earlier studies (Gupta *et. al.*, 2010), we trained and tested different state-of-art ensemble and meta classification methods for identification of Pol-II enriched promoter and Pol-II enriched non-promoter sequences, each of length 500 bp. The classification models were trained and tested on a bench-mark dataset, using a set of 39 different feature variables that are based on chromatin modification signatures and various DNA sequence features. The best performing model was implemented in a promoter prediction algorithm that was applied on seven published ChIP-seq Pol-II datasets to provide genome wide annotation of mouse gene promoters.

Application of statistical methods successfully used in the field of linguistics: Back in 1990s, assuming nucleic acid sequences as words over the alphabet of nucleotides, computational biologists have applied statistical methods borrowed from the field of linguistics to show that the DNA indeed has all the features of a human language, ranging from alphabets and lexicons to grammar and phonetics (Brendel and Busse, 1984; Head, 1987; Searls, 1992; Ji, 1999). Searls, in his pioneering work (Searls, 2002), demonstrated that the sequential interpretation of DNA sequence to structure to function could be matched to the hierarchical model of the human language. Furthermore, it was shown that the noncoding DNA sequences were more similar to natural languages than the coding regions (Mantegna *et. al.*, 1994). One crucial feature of any language is the semantic dependency on the verbal context, which refers to the surrounding text of a particular word or sentence of interest. For example, the correct

meaning polysemous words, which spell the same but are semantically different, can only be inferred by the clues in the context (Peters *et. al.*, 2018). Similarly, in non-coding promoter region, a Transcription Factor Binding Site (TFBS) can be target of different transcription factors that belong to the same family of proteins. For example, p53 family of proteins (p53, p63, p73) share the same central DNA-binding domain structure, which binds as a tetramer to consensus response elements (RE) consisting of two decameric palindromic half-site sequences (Ethayathulla *et. al.*, 2013; Kearns *et. al.*, 2016). Many studies have suggested that p53 family members recognize the same RE sequences while exhibiting selectivity in actual binding (Osada *et. al.*, 2005; Lokshin *et. al.*, 2007; Schavolt and Pietenpol, 2007). In this case, the conserved binding site sequence is polysemous in that the semantic meaning (whether it is to be bound by any isoform of p53, p63 or p73) varies and is hard to determine from the motif sequence alone without taking the context into consideration.

Application of traditional machine-learning methods: Computational approaches that combine seemingly disparate experimental data have been successful in developing accurate classification models. For example, RandomForest (Breiman, 2001) based algorithms have been receiving increased attention in the data-science field as a means of variable selection in many classification tasks in computational biology, including the selection of a subset of genetic markers (Lunetta *et. al.*, 2004; Bureau *et. al.*, 2005) and genes in microarray data analysis (Cutler and Stevens, 2006; Pang *et. al.*, 2006) relevant for the prediction of a certain disease.

We have earlier used an integrative modeling approach that combines CART (Breiman, 1984) and RandomForest to classify different Estrogen Receptor alpha (ERα) responsive promoters (Cheng *et. al.*, 2006) and SMAD target promoters (Qin *et. al.*, 2009) with reasonably good classification accuracy and reduced instability (Qin *et. al.*, 2009). Although the main goal in classification is to build a model with minimal mis-classification error in cross-validation, in these applications we were equally interested in identifying TFBSs as highly important discriminating variables, between different groups of promoter sequences. One of the main goals of our analyses is to select potential TFBS from a large feature space (>1,000) in order to build binary classifiers. RandomForest algorithm generates internal estimates of the decrease in the classifier's overall accuracy if that particular variable was not used in building the classifier. Thus, variables (TFBSs in this case) with larger importance measures can be deemed to have more power in discriminating different groups.

This integrative microarray data-analyses and statistical modeling approach facilitated the prediction of which proteins work with estrogen to contribute to breast cancer development. The computational predictions in this study indicated that the interaction of estrogen with one of seven different partner proteins determines whether the gene is activated or suppressed in breast cancer cells. This was a noteworthy machine-learning methodology breakthrough because it allowed integrative analysis of big datasets, often consisting of expression, chromatin landscaping data and TF-binding information for thousands of genomic loci, for predicting groups of TFBS (known as cis-regulatory modules – CRMs), which then could be validated by more traditional experimental biology techniques. In addition, this novel computational methodology has been applied in several other subsequent studies within our group and outside for integrative analysis of transcriptome and genome data.

Application of deep-learning methods: In recent years, many computational tools have been developed by successfully applying deep learning techniques on genomic sequence data to

study the individual aspects of *cis*-regulatory landscapes, including DNA-protein interactions (Alipanahi *et. al.*, 2015), chromatin accessibility (Kelley *et. al.*, 2016), non-coding variants (Zhou and Troyanskaya, 2015). Most methods adopted Convolutional Neural Network (CNN)-based architecture (Zou *et. al.*, 2019). Other tools focus on the sequential characteristic of DNA and attempt to capture the state dependency using Recurrent Neural Network (RNN)-based models, such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho *et. al.*, 2014) networks. Several hybrid methods were also proposed to integrate the two model architectures (Hassanzadeh and Wang, 2016; Quang and Xie, 2016; Shen *et. al.*, 2018). To better model gene regulatory elements in non-coding DNA, an ideal computational model should (i) globally take all the contextual information into account to understand polysemous CREs; (ii) develop generic understanding transferable to various tasks; (iii) generalize well when labeled data is limited. However, both CNN and RNN architectures fail to satisfy these requirements (Bengio *et. al.*, 2013; LeCun *et. al.*, 2015). CNN is usually unable to capture long-range semantic dependency due to the limited filter size. RNN models (LSTM, GRU), although able to learn long-term dependency, greatly suffer from vanishing gradient and low-efficiency problem when it sequentially processes all past states and compresses contextual information into a bottleneck with long input sequences. In addition, a significant challenge is that most existing prediction models require massive amounts of labeled data, resulting in limited performance and applicability in data-scarce scenarios.

Both CNN and RNN network architectures have their intrinsic limitations in terms of understanding whole genome as a language. Gene regulatory components separated by hundreds or even thousands of nucleotides are often found to coordinate together, which suggests the existence of distant semantic relationship within contexts. However, CNN is usually unable to capture such long-range semantic dependency, as its capability to extract features is limited by the size of the sliding kernels. In other words, CNN can effectively learn the local features but is not efficient at understanding the sequential relationship between these segments. LSTM and GRU, in contrast, specialize in modeling time-series data and are effective at capturing the dependency. Nonetheless, a major limitation of RNN is the difficulty for parallelized computation, since sequential information at time t is intrinsically dependent upon completion of all past states. This greatly limits the efficiency of training large RNN-based models with many blocks/layers. A combination of CNN and RNN still cannot bypass the sequential processing problem of RNN and is, therefore, also suboptimal when the model size gets large.

In essence, the goal for any deep learning model is to find a good distributed representation of input that is best suited for the downstream tasks (e.g. classification) (Bengio *et. al.*, 2013; LeCun *et. al.*, 2015). In NLP, such representation is often called embeddings, which are low-dimensional numeric vector representations of words or sentences that allows for arithmetic operations (Mikolov *et. al.*, 2013b). Ng (Ng, 2017) developed dna2vec as an initial method in 2017 based on the popular word2vec model (Mikolov *et. al.*, 2013a), which computes word embeddings of variable-length k-mers of input DNA sequence with a skip-gram model architecture. Essentially, each k-length fragment of input sequence will be considered as a 'word' and converted into a dense vector representation, by predicting the words immediately surrounding it. However, one major drawback of such embedding obtained is that the resulting representation will be context-independent (Peters *et. al.*, 2018). Two subsequences "ATGCCA" will always have same vector representation, whereas in reality they may mean different things. Same limitation applies to representations learned by CNN models, which are also not contextual (Yan and Guo, 2019). Even for RNN models, it becomes difficult

for the model to compress information from all previous states into a "bottleneck" when the input sequence gets long (Bahdanau *et. al.*, 2014).

To more properly train a model that can develop contextual embedding, attention mechanism was proposed so that the model should not put equal weights on everything it learned (Bahdanau *et. al.*, 2014; Yang *et. al.*, 2016). Instead, it should learn to "attend to" the important parts by properly forming a context vector, which is essentially a weighted sum of all hidden states from the encoder (Bahdanau *et. al.*, 2014). The weights, or attention, measures the alignment between the output and every encoder-hidden state. After its initial success in machine translation, attention mechanism became widely applied in various types of tasks, including image captioning (Xu *et. al.*, 2015), speech recognition (Chorowski *et. al.*, 2015) and others. Recently, attention-based networks are also applied in bioinformatics for predicting enhancer-promoter interaction (Mao *et. al.*, 2017; Hong *et. al.*, 2020). However, most of attention networks developed typically rely on use of a RNN-based architecture, which is subjected to the drawback mentioned above (Vaswani *et. al.*, 2017).

In order to address the limitations listed above, a new type of model that simultaneously possesses the strengths of both CNN and RNN and is capable of developing contextual embedding via attention mechanisms is necessary.

Adaptation of BERT for genome sequence prediction tasks: The advent of the Bidirectional Encoder Representation Transformer (BERT) model lead the NLP research to a new era by introducing a paradigm of pre-training and fine-tuning. Major technical innovation of BERT is bidirectional training of Transformer model, a popular attention model to language modelling (Vaswani *et. al.*, 2017). This is in contrast to previous methods, which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. It was shown that bi-directionally trained language models superior performance by capturing deeper understanding of language context and flow than single-direction language models (Devlin *et. al.*, 2018). BERT enables effective use of unlabeled text by proposing novel pre-training tasks; masked language modeling and next sentence prediction. The tasks guide a model to learn contextualized representations of words and relationship between sentences. Models are first pre-trained on a massive amount of unlabeled data to learn the general rules and relationships; and then fine-tuned on task-specific labeled data to learn to perform specific classification tasks.

Based on promising results in NLP research, we hypothesized that pre-trained transformer-based neural network model offer a promising, and yet not fully explored, deep learning approach for a variety of sequence prediction tasks in the analysis of non-coding DNA. To investigate this, we recently developed a pre-trained bidirectional encoder representation, named DNABERT, for global interpretation of genomic sequences based on up and downstream nucleotide contexts (Ji *et. al.*, 2021). DNABERT out-performed most widely used programs for genome-wide regulatory elements prediction in accuracy and efficiency. Single pre-trained DNABERT model could simultaneously achieve state-of-the-art performance on prediction of promoters, splice sites, and transcription factor binding sites, after easy fine-tuning using small task-specific labelled data. Further, DNABERT enabled direct visualization of nucleotide-level importance and semantic relationship within input sequences for better interpretability and accurate identification of conserved sequence motifs and functional genetic variant candidates. The pre-trained DNABERT with human genome could also be readily applied to other organisms with exceptional performance.

4. Conclusions

Most of the traditional bioinformatics tools develop the understanding of DNA from scratch with task-specific data. As the deep learning models become gradually deeper and wider, their demand for data is getting much more intense. Thus, simply relying on labeled data is very likely to result in poor performance when dataset size is small. In contrast, BERT-style *pre-train—fine-tune* scheme ingeniously utilizes the massive amount of unlabeled data to gain such understanding without the need for any human guidance, while such understanding it obtained is easily transferable to various downstream tasks. Therefore, the model can still achieve exceptional performance in data-scarce scenarios. Second, in comparison to CNN architecture, which only captures local context, Transformer globally capture contextual information from the entire input sequence by taking all the representations from the last layer as input and performing self-attention on them. With the self-attention mechanism, Transformer is not only straightforwardly parallelizable, but also effectively overcomes the gradient vanishing problem that RNN-based architectures usually meet. Therefore, BERT style modeling is expected to lead to many biological breakthroughs through general understanding of DNA by correctly capturing the hidden syntax

In the design of the above mentioned and other successful computational prediction programs, combining state-of-the-art statistical pattern recognition methods with computational systems design is essential. Additionally, expertise in data-curation and computer programing is key in the development of successful algorithms and bioinformatics software for solving these and several other challenging problems in mammalian genomics.

Acknowledgements

This work was supported by the National Library of Medicine of the NIH [R01LM011297 to RD].

References

Alipanahi, B., Delong, A., Weirauch, M. T. and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, **33**, 831-838.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:14090473*.

Bengio, Y., Courville, A. and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798-1828.

Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont, California.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5-32.

Brendel, V. and Busse, H. G. (1984). Genome structure described by formal languages. *Nucleic Acids Research*, **12**, 2561-2568.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, **28**, 171-182.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78-94.

Cheng, A. S., Jin, V. X., Fan, M., Smith, L. T., Liyanarachchi, S., Yan, P. S., Leu, Y. W., Chan, M. W., Plass, C., Nephew, K. P. et al. (2006). Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Molecular Cell*, **21**, 393-404.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:14061078*.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pp. 577-585.

Cutler, A. and Stevens, J. R. (2006). Random forests for microarrays. *Methods Enzymology*, **411**, 422-432.

Davuluri, R. V. (2003). Application of FirstEF to find promoters and first exons in the human genome. In *Current Protocols in Bioinformatics*, Vol **4.7** (ed. A Baxevanis). John Wiley & Sons, New York, NY.

Davuluri, R. V., Grosse, I. and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nature Genetics*, **29**, 412-417.

Davuluri, R. V. and Zhang, M. Q. (2003). Computer software to find genes in plant genomic DNA. *Methods in Molecular Biology*, **236**, 87-108.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*.

Dunham, I. Kundaje, A. Aldred, S. F. Collins, P. J. Davis, C. Doyle, F. Epstein, C. B. Frietze, S. Harrow, J. Kaul, R. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

Ethayathulla, A. S., Nguyen, H. T. and Viadiu, H. (2013). Crystal structures of the DNA-binding domain tetramer of the p53 tumor suppressor family member p73 bound to different full-site response elements. *Journal of Biological Chemistry*, **288**, 4744-4754.

Gupta, R., Wikramasinghe, P., Bhattacharyya, A., Perez, F. A., Pal, S. and Davuluri, R. V. (2010). Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics* **11 Suppl 1**: S65.

Hassanzadeh, H. R. and Wang, M. D. (2016). DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 178-183. IEEE.

Head, T. (1987). Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *Bulletin of Mathematical Biology*, **49**, 737-759.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computing*, **9**, 1735-1780.

Hong, Z., Zeng, X., Wei, L. and Liu, X. (2020). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics*, **36**, 1037-1043.

Ji, S. (1999). The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Annals of the New York Academy of Sciences-Paper Edition*, **870**, 411-417.

Ji, Y., Zhou, Z., Liu, H. and Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* doi:10.1093/bioinformatics/btab083.

Kearns, S., Lurz, R., Orlova, E. V. and Okorokov, A. L. (2016). Two p53 tetramers bind one consensus DNA response element. *Nucleic Acids Research*, **44**, 6185-6199.

Kelley, D. R., Snoek, J. and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, **26**, 990-999.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol*, **4**, 134-142.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521**, 436-444.

Lokshin, M., Li, Y., Gaiddon, C. and Prives, C. (2007). p53 and p73 display common and distinct requirements for sequence specific binding to DNA. *Nucleic Acids Research*, **35**, 340-352.

Lunetta, K. L., Hayward, L. B., Segal, J. and Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, **5**, 32.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M. and Stanley, H. E. (1994). Linguistic features of noncoding DNA sequences. *Physical Review Letters*, **73**, 3169-3172.

Mao, W., Kostka, D. and Chikina, M. (2017). Modeling enhancer-promoter interactions with attention-based neural networks. *bioRxiv*: 219667.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111-3119.

Ng, P. (2017). dna2vec: Consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:170106279*.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. and O'Neal, C. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, **53**, 1161-1168.

Osada, M., Park, H. L., Nagakawa, Y., Yamashita, K., Fomenkov, A., Kim, M. S., Wu, G., Nomoto, S., Trink, B. and Sidransky, D. (2005). Differential recognition of response elements determines target gene specificity for p53 and p63. *Molecular Cell Biology*, **25**, 6077-6089.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E. and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028-2036.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:180205365*.

Portin, P. (2014). The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. *Journal of Genetics*, **93**, 293-302.

Qin, H., Chan, M. W., Liyanarachchi, S., Balch, C., Potter, D., Souriraj, I. J., Cheng, A. S., Agosto-Perez, F. J., Nikonova, E. V., Yan, P. S. et al. (2009). An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. *BMC System Biology*, **3**, 73.

Quang, D. and Xie, X. H. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, **44**.

Schavolt, K. L. and Pietenpol, J. A. (2007). p53 and Delta Np63 alpha differentially bind and regulate target genes involved in cell cycle arrest, DNA repair and apoptosis. *Oncogene*, **26**, 6125-6132.

Searls, D. B. (1992). The linguistics of DNA. *American Scientist*, **80**, 579-591.

Searls, D. B. (2002). The language of genes. *Nature*, **420**, 211-217.

Shen, Z., Bao, W. and Huang, D.-S. (2018). Recurrent neural network for predicting transcription factor binding sites. *Scientific Reports*, **8**, 1-10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998-6008.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048-2057.

Xu, Y., Mural, R., Shah, M. and Uberbacher, E. (1994). Recognizing exons in genomic sequence using GRAIL II. *Genetic Engineering*, *(N Y)* **16**, 241-253.

Yan, D. F. and Guo, S. Y. (2019). Leveraging Contextual Sentences for Text Classification by Using a Neural Attention Model. *Computational Intelligence and Neuroscience*, **2019**.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480-1489.

Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Review Genetics*, **3**, 698-709.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, **12**, 931-934.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A. and Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, **51**, 12-18.