Multi-Headed Attention Networks:

A Novel Approach to Biological Sequence Analysis

Harissh Ragav Dhamodaran, Sadhana Thirumangai Kalidoss,
Gauri Lekshmi Sathya, Nimisha M Iyer, Shravan Sunil, Karthik M

Abstract

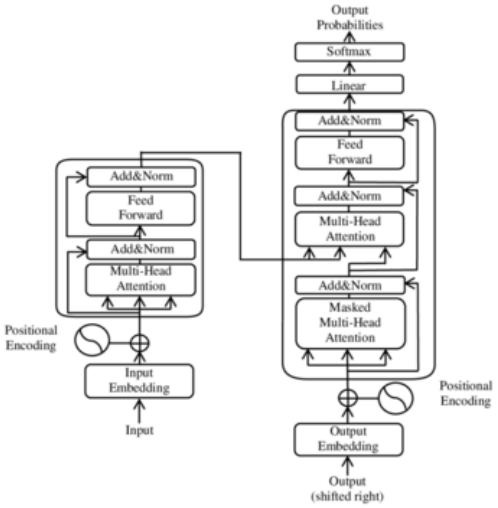
The ongoing evolution of viruses poses a significant challenge to public health, necessitating the development of predictive models to predict genetic mutations in virus DNA. This paper presents a novel approach to predicting mutations in virus DNA sequences through the utilization of Multiheaded Attention, a deep learning mechanism originally designed for natural language processing tasks. By adapting this architecture to the genomics domain, we aim to enhance our ability to forecast viral mutation events accurately and use it to our advantage.

---

1. Introduction

Transformers are a type of neural network architecture which was developed for the problem of sequence transduction, or neural machine translation. This means that any task which involves the transformation of an input sequence into an output sequence can be solved using this model. Such tasks include speech recognition, text-to-speech transformation, language translation, etc. Until now, Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) have been used for such tasks due to their properties. Albeit giving us accurate results, RNNs become very ineffective when the gap between the relevant information and the point where it is needed becomes very large. That is due to the fact that the information is passed at each step and the longer the chain is, the more probable the information is lost along the chain. CNNs also do not help with the problem of understanding dependencies when translating statements. To tackle this issue, we use the technique of Attention. Transformers, in essence, are a combination of CNNs and Attention. Specifically, Transformers utilize self-attention to boost the speed of translation from one sequence to another. As it is a sequence transducer, we have utilized pre-existing genome codes of viruses and their mutants for our AI to generate possible mutations that could occur in the future. With the mutated genome sequences in hand, we can extrapolate critical data regarding virus genomes.

2. Model Architecture



The Transformer - model architecture. [1]

Most competitive neural sequence transduction models have an encoder-decoder structure. We employ self-attention to train our model. Self-attention, also known as intra-attention, is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence. It has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations. In the encoder, we begin by transforming our data from a natural language sequence into matrices which the neural network can understand. The standard transformer model has six identical encoding layers and six identical decoding layers.

2.1. Encoder

Each layer in the encoder has two sub-layers, one being a multi-head self-attention mechanism whereas the other is a position-wise fully connected feed-forward network. We add a residual connection from around each of the sub-layers and follow it up with layer normalization. The output of each sub-layer is LayerNorm(x + Sublayer(x)). Where Sublayer(x) is the function implemented by the sub-layer itself. To smooth the way for our residual connections, all sub-layers produce outputs in the dimensions of 512.

2.2. Decoder

The decoder is also composed of six layers. With the two sublayers present, the decoder also has another third sublayer, which performs multi-head attention over the output of the encoder stack. Just like the encoder, we use residual connections and follow it up with layer normalization. Additionally, we modify the self-attention in the sublayer in the decoder stack to prevent positions from attending to subsequent positions. In layman's terms, it prevents our AI from actively 'knowing' what it is going to generate next. This is known as masking. The masking operation is a filter matrix which assigns a value of negative infinity to all the codons that will be generated in the future, thus masking the future codons. Once the masked-attention filter is passed through the softmax layer, all the values of negative infinity get zeroed out. Therefore, while predicting the next codon, the model only

pays attention to the codons that have already been generated. When combined with offsetting the output embeddings by one position, it is ensured that the predictions for a specific position can only depend on the outputs that have been generated before it.

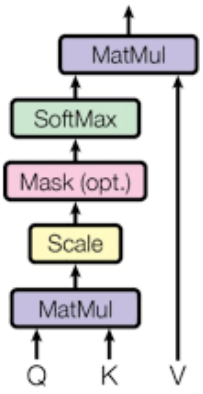
2.3. Positional Encoding

In our specific example of genome sequencing, we form a vocabulary dictionary out of all the genome sequences that we have used as input. This data is then assigned a numeric index for every 3 nucleotides as 3 nucleotides make a codon, coding for what protein must be synthesized by the organism. Hence, the input of the transformer is not the genetic sequence itself, but the corresponding indices. Tokenizers take care of this for us. These indices are denoted with (x_1, ..., x_n). This is known as processing the input. The transformer model does not have any kind of recurrence or convolution, hence, we need a way to retain the information about the relative or absolute position of the tokens in sequence. Hence, we add positional encodings to the input embeddings at the bottom of the encoder and decoder stacks. Position encoding is a fixed-size vector representation which stores the relative positions of the tokens within a target sequence. We take the indices of the corresponding input and attach a position vector to it. The embedding layer selects the embedding corresponding to the input statement, and then further passes it on. By utilizing positional encoding, the AI is enabled to know the position of the codon in the given sequence, which reduces the time taken to train the neural network.

2.4. Scaled dot-product attention

The transformer building blocks are scaled dot-product attention units. For each attention unit, the transformer model learns three weight matrices: the query weights W_Q, the key weights W_K, and the value weights W_V. For each token i, the input token representation x_i is multiplied with each of the three weight matrices to produce a query vector $q_i = x_i W_Q$, a key vector $k_i = x_i W_K$, and a value vector $v_i = x_i W_V$. Attention weights are calculated using the query and key vectors: the attention weight a_{ij} from token i to token j is the dot product between q_i and k_j. The attention weights are divided by the square root of the dimension of the key vectors, $\sqrt{d}_k$ which stabilizes gradients during training, then passed through a softmax which normalizes the weights. The attention calculation for all tokens can be expressed as one large matrix calculation using the softmax function. Hence, the final equation to represent attention is

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



Scalar Dot-Product Attention.[2]

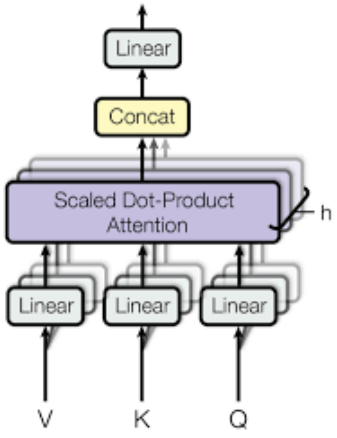
2.5. Multi-head attention

One set of (W_Q, W_K, W_V) matrices is called an attention head. Each layer in the transformer model has multiple attention heads. The computations for each attention head can be done in parallel, which allows for faster processing. The outputs for the attention layer are then concatenated to pass into the feed-forward neural network layers. If the multiple attention heads are indexed by i, then we have the following equation.

$$MultiHead(Q,K,V)=Concat(head_1,...,head_n)W^O$$

$$head_i=Attention(XW_i^Q, XW_i^K, XW_i^V)$$

Where the matrix X is the concatenation of word embeddings, the matrices W_i^Q, W_i^K, W_i^V are 'projection matrices', owned by the individual attention head i, and W^O is the final projection matrix owned by the whole multi-headed attention head.



Multi-Head Attention.[2]

3. Applications

Once we have our sequenced genome, we can apply it in many different fields. At the heart of our project, we aim to use the code for drug development. Genomics is a topic that has gained traction over the years; we handle many different methodologies and new ideas from our long string of codons. A similar use of ML to predict virus mutations has been used to predict the mutations of the Influenza-A virus and recurrent mutations in cancer. This testifies to the utility and credibility of artificial intelligence in biotech and medical fields.

3.1. Genomics

Genome analysis is the prerequisite for all the other applications. It is the process of studying and interpreting the genetic information contained within an organism's genome. The genome is the absolute set of instructions for the creation of a biologically live organism.

Genome sequence data helps in the development of genomics. It helps us tackle problems related to the structures of macromolecules that are encoded in the genome. We can predict the mutations that could occur in the virus. We can use that data to understand the changes that could occur in the amino acids of the virus and the implications it has.

The key aspect of genome analysis is known as Annotation. It is performed to locate and understand the function of each part of the genome. Annotation is a crucial step in gene analysis as it allows the researchers to translate the raw genome sequence data into something that they can use to study and answer questions related to the biology of the organism. In annotation, we identify and label the various parts of the genome sequence as genes, repetitive sequences, promoters and regulatory regions, and non-coding RNAs. The most crucial among these is to identify the parts of the genome which code for proteins, called genes. Computational tools and algorithms are used to decipher the locations of genes.

Functional annotation specifically deals with the identification of the function of a particular gene. We compare the identified genes with a database of known genes to assign functions. Other than purely computational methods. We also use experimental data to identify gene functions. This is better known as Comparative Annotation. This is of significance in studying closely related variants of viruses and other pathogens. A study in this approach will help us pinpoint the exact mutation which causes the new trait in a variant.

Our AI can not only predict mutations in viruses but also estimate the bases in a genome sequence. When we sequence gene codes from a physical sample, occasionally we're unable to identify all of the bases in a sequence. Hence, in place of a nucleotide, we see the letter 'N', marking an ambiguous nucleotide. This is due to the physical limitations in our machinery. Our AI can analyze the genome sequence and fill in the nucleotides which should be in the

placeholders, helping us attain clarity regarding gene sequencing from physical samples.

We may also analyze and spot those regions of the DNA that are highly conserved and thereby predict that these segments serve some essential function to the organism and devise our vaccines and drugs on its basis.

3.2. Preventive Action

Our present method of preparing for a pandemic involves waiting for a new variant to emerge before we sequence its genome and then flag it as a potential variant of concern (VOC). Variants of concern are those that have increased transmissibility, severe disease-causing capabilities, and more immunity towards antibodies developed through previous infections. Therefore, they have higher chances of evading diagnosis through existing methods. Using our AI model, we could predict the mutations that a virus, which could potentially cause an outbreak, undergoes beforehand. Consequently, we have a better chance of being prepared for such an outbreak. Our AI acts as a tool that is used to determine whether the diagnostic tools and the vaccines and treatments that were tailored for the earlier version of the virus would now work on the mutated virus. Therefore, the model has immense applications in supporting our healthcare system. Mutations can be mapped to known epitome sites and regions of the genome that are known to be involved in antibody escape to further analyze its impacts on resistance. The implementation of an early warning system that analyzes the data over months and can flag the rapidly mutating variants allows us to gain a proactive stance on biological threats.

3.3. Reverse Vaccinology

The traditional vaccine development technology involves injecting the inactivated or weakened pathogens into the body to stimulate immunity. Since we require vaccines for the upcoming mutations of our selected virus, we use a technology known as reverse vaccinology. It incorporates designing vaccines using sequenced genomes of pathogens. Reverse vaccinology takes a neoteric route by leveraging bioinformatics and computational analysis of pathogen genomes to identify potential vaccine targets. The process of development of vaccines using this method involves cloning and the expression of all proteins in an organism's genome sequence. Reverse vaccinology has already been used in the development of vaccines for malaria, anthrax etc.

3.4. Antiviral Drug Development

The current antiviral drugs and antiviral antibodies, which are specific at protein levels, have encountered difficulties in being effective due to the rapid evolution of mutant viral strains resulting in drug resistance. Therefore, degrading viral genomes for the development of antiviral

drugs is an innovative approach providing us with greater efficacy. A specialized branch

known as functional genomics, which takes up the study of how genes contribute to different metabolic pathways, can be applied to study the gene sequence of the mutated virus and pave the way for progress in antiviral drugs.

4. Conclusion

By developing a neural network to predict mutations of a virus, we contribute to genomics, which provides us with an incredible opportunity in vaccine development where the traditional methods have failed. Keeping the evolution of structural genomics in mind, genes now have the potential to be a vital tool in the development of drugs and vaccines in the future. Additionally, we are able to extrapolate data regarding the structure and functions of gene segments.

This software enables us to make vital inferences about genome sequencing and its applications.

References

[1] Y. Jia, "Attention Mechanism in Machine Translation," *Journal of Physics: Conference Series*, vol. 1314, p. 012186, Oct. 2019, doi: https://doi.org/10.1088/1742-6596/1314/1/012186

[2] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv.org*, Dec. 05, 2017. doi: https://doi.org/10.48550/arXiv.1706.03762. Available: https://arxiv.org/abs/1706.03762

[3] M. Elez, L. Robert, and I. Matic, "Method for Detecting and Studying Genome-Wide Mutations in Single Living Cells in Real Time," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1736, pp. 29–39, 2018, doi: https://doi.org/10.1007/978-1-4939-7638-6_3. Available: https://pubmed.ncbi.nlm.nih.gov/29322456/

[4] C. J. Houldcroft, M. A. Beale, and J. Breuer, "Clinical and biological insights from viral genome sequencing," *Nature Reviews Microbiology*, vol. 15, no. 3, pp. 183–192, Mar. 2017, doi: https://doi.org/10.1038/nrmicro.2016.182. Available: http://www.nature.com/articles/nrmicro.2016.182. [Accessed: May 21, 2020]

[5] T. G. Magaldi *et al.*, "Mutations in the GM1 Binding Site of Simian Virus 40 VP1 Alter Receptor Usage and Cell Tropism," *Journal of Virology*, vol. 86, no. 13, pp. 7028–7042, Apr. 2012, doi: https://doi.org/10.1128/jvi.00371-12

[6] J. Goulding, "Virus replication | British Society for Immunology," *www.immunology.org*. Available: https://www.immunology.org/public-information/bitesized-immunology/pathogens-disease/virus-replication

[7] S. Khanal, "Mechanism of Action of Antiviral Drugs," *Microbe Online*, Oct. 04, 2022. Available: https://microbeonline.com/mechanism-of-action-of-antiviral-drugs/

[8] S. Kausar *et al.*, "A review: Mechanism of action of antiviral drugs," *International Journal of Immunopathology and Pharmacology*, vol. 35, p. 205873842110026, Jan. 2021, doi: https://doi.org/10.1177/20587384211002621. Available: https://journals.sagepub.com/doi/full/10.1177/20587384211002621

[9] M. C. Maher *et al.*, "Predicting the mutational drivers of future SARS-CoV-2 variants of concern," *Science Translational Medicine*, vol. 14, no. 633, Feb. 2022, doi: https://doi.org/10.1126/scitranslmed.abk3445

[10] M. A. Salama, A. E. Hassanien, and A. Mostafa, “The prediction of virus mutation using neural networks and rough set techniques,” *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2016, no. 1, May 2016, doi: https://doi.org/10.1186/s13637-016-0042-0

[11] B. Saldivar-Espinoza *et al.*, “Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks,” *International Journal of Molecular Sciences*, vol. 23, no. 23, p. 14683, Nov. 2022, doi: https://doi.org/10.3390/ijms232314683. Available: https://pubmed.ncbi.nlm.nih.gov/36499005/. [Accessed: Aug. 06, 2023]

[12] “New AI can predict virus mutations and help create more effective treatments and vaccines,” *Waterloo News*, Oct. 25, 2021. Available: https://uwaterloo.ca/news/media/new-ai-can-predict-virus-mutations-and-help-create-more

[13] “COVID-19 Genomes,” *www.kaggle.com*. Available: https://www.kaggle.com/datasets/tunguz/covid19-genomes?resource=download. [Accessed: Sep. 20, 2023]

[14] I. Delany, R. Rappuoli, and K. L. Seib, “Vaccines, Reverse Vaccinology, and Bacterial Pathogenesis,” *Cold Spring Harbor Perspectives in Medicine*, vol. 3, no. 5, pp. a012476–a012476, May 2013, doi: https://doi.org/10.1101/cshperspect.a012476