Applying NLP Models for the Prediction of siRNA Sequences Based on DNA

*Mica Haney, Camrie Johnson*

# Abstract

Small interfering RNAs (siRNAs) are a class of RNAs ranging from 20-30 nucleotides in length with the ability to silence gene expression with specificity and efficiency. Interest has grown substantially in siRNAs as a treatment for many diseases given their broad range in application. Many computational models have been developed to predict successful siRNA sequences. Even so, no current models in this field have utilized natural language processing (NLP) frameworks. Three models were created for this study including Naïve-Bayes, long-term short memory (LTSM), and a transformer. All three were used to generate siRNA sequences for 14 different genes. Results were validated based on three metrics: Hamming distance, BLUE, and ROUGE. Our results suggest that the Transformer model outperformed other models explored in this study with a ROUGE and BLUE score exceeding 0.7. Overall, results were modest but the trends within the results regarding the Transformer were promising.

# 1 Introduction

RNA interference (RNAi) is a biological mechanism in which double stranded RNA silences complementary strands of messenger RNA (mRNA) (Dana et al., 2017). There are two classes of noncoding RNAs that play a role in RNAi which are microRNAs (miRNA) and small interfering RNAs (siRNA). siRNAs are around 20-30 nucleotides in length and inhibit the expression of one specific mRNA sequence. Conversely, miRNAs can inhibit expression of multiple mRNAs (Lam et al., 2015). This project will focus on siRNAs.

Both classes undergo similar mechanisms upon entering the cell (Lam et al., 2015). The siRNA enters the cell in a double stranded form and is loaded onto an RNA-induced silencing complex (RISC). Argonaute 2 (AGO2) is an endonuclease within RSIC that cleaves the sense, or passenger, strand leaving only the antisense, or guide, strand within the complex. This guide strand directs RISC to the target mRNA which is then cleaved by AGO2 (Figure 1). This in turn silences the gene by preventing the mRNA from being translated into a protein.

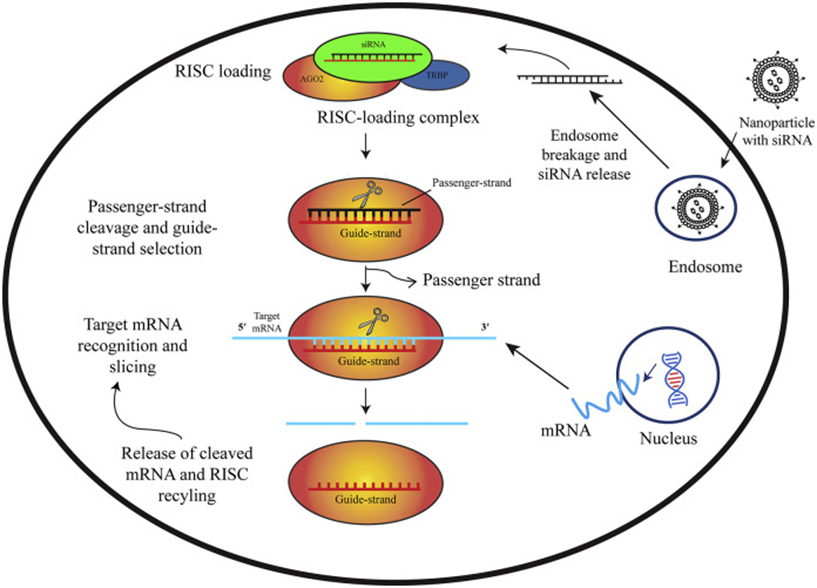


Figure 1: Mechanism of siRNA silencing (Chowdhury et al., 2021)

This technology has the potential to treat any genetic ailment. siRNA silencing has been applied to diseases such as COVID-19, cancers, and brain diseases(Chowdhury et al., 2021; Singh et al., 2018; Zheng et al., 2018). Though there is a broad range in treatment opportunities, there are challenges in designing an effective siRNA therapy. Occasionally a design may induce off-target effects by interacting with near perfect mRNA complementary sequences or degrade before reaching its target (Ghosal et al., 2012).

Computational methods can be employed to determine the optimal siRNA sequences for a target gene. Genome wide siRNA libraries have been developed using artificial neural networks (Huesken et al., 2005). Some models predict the efficacy of modified siRNAs such as cholesterol conjugated siRNAs (Shmushkovich et al., 2018). However, there are no models making use of the natural language processing (NLP) frameworks within this area. NLP is a subfield of artificial intelligence and is typically used in processing human language (Zhou et al., 2020). Though this method may seem unconventional, one could view nucleotide sequences as words given that each nucleotide base is denoted as a letter (A, C, T, G and U).

The aim of this project is to develop an NLP model to predict siRNA sequences. The data used in this study was pulled from another study in which a neural network was developed to predict siRNA sequences based on thermodynamic properties (Shabalina et al., 2006). The parameters used in the previous study are also used in our NLP model. The first of the three parameters include the thermodynamic Gibbs free energy difference (ΔG) between the 5’ ends of the sense and antisense strands of siRNA. The second parameter is the dinucleotide content indexes which for the purpose of this study were normalized between 0 and 1. The last parameter was a full position dependent consensus, which accounts for significantly preferred and avoided nucleotides in all sequence positions.

# 2 Methods

## 2.1 Data

### 2.1.1 Data Description and Collection

The siRNA sequences and all associated information aside from the target gene sequence was collected from (Shabalina et al., 2006) with , while the target gene sequences were collected from (National Center for Biotechnology Information, 2023). The initial dataset contained 650 records. However, only 608 were kept because some of the genes’ DNA sequences could not be retrieved. The data was collected using a web scraper implemented using the Python libraries selenium, urllib, and BeautifulSoup.

We used the following features from the dataset from (Shabalina et al., 2006):

* Target seq: The GeneBank sequence accession number of the gene the siRNA targets.
* Start and End: The start and end position of the siRNA binding site in the DNA sequence.
* Sequence: The siRNA sequence.
* G and U: The counts of the G and U content in the siRNA sequence, respectively.
* Diff\_5-3: Energy difference (ΔG) between 5’ and 3’ ends.
* Cons\_Sum: Consensus summary of position dependent nucleotides.
* Activity: The amount of silencing that the siRNA performs on its target as a percentage.

The other features in the dataset were determined to be not relevant to this project’s task on a biological level and therefore not used.

### 2.1.2 Preprocessing

The data was first split into training and testing sets at an 80 to 20 ratio. The data split into three datasets. The sequences dataset contained all of the target DNA sequences, the siRNA sequences, and the locations of where the siRNA sequences bound to in the DNA sequence. The activity dataset contained the same features as the sequences dataset, however all rows where the activity feature was less than 70% were dropped. The numerical dataset contained all of the rows and features in the sequences dataset but included 6 additional numerical features: Dif\_5-3, A, C, G, U content, Cons\_Sum. The numerical features were normalized to a range of 0 to 1.

For the Naive Bayes classifier, the dataset was preprocessed into sliding windows. For each DNA/siRNA pair the siRNA was mapped to the binding position. For each position in the siRNA, the previous 5 characters in the DNA sequence were collected and labeled with a target class of the character in the siRNA at that position. The sliding windows were then one-hot encoded.

For the LSTM and Transformer for each record the section of the DNA that the siRNA binds to was taken as a string. Each DNA subsequence and siRNA sequence had a start and end character added onto the start and end of each sequence, respectively. The sequences were then label encoded.

## 2.2 Models

The data was used to train three types of machine learning models: a chained Naive Bayes classifier, a LSTM, and a Transformer.

The Naive Bayes was implemented using scikit-learn’s CategoricalNB class. Training was done over the windowed data described in Section 2.1. Sequence prediction was done by predicting individual characters for each window of a sequence in order and concatenating the predictions into a string.

The LSTM was implemented using Tensorflow’s predefined model and layer classes. There was a 256 unit embedding layer followed by two bidirectional LSTM layers, the first with 256 hidden units and the second with 128 hidden units. Both bidirectional LSTM layers were followed by a 0.1 dropout layer. A last dense layer for prediction was added. All of the LSTM models were compiled with Tensorflow’s default Adam optimizer and sparse categorical cross-entropy loss.

The Transformer was implemented using Tensorflow’s implementation in (Neural machine translation) without the custom learning schedule and modification to the data handling steps necessary to train on the project’s data. All Transformer models were built with 2 encoder/decoder layers, each with 128 hidden units in the embedding layer and 256 units everywhere else. Each attention layer had 4 heads. The models were compiled with Tensorflow’s default Adam optimizer and a masked sparse categorical cross-entropy loss and masked accuracy.

All models were trained using 5-fold cross-validation on each of the three datasets and tested on the held-out test sets. All models were trained for 10 epochs (per fold) with batch sizes of 20.

# 3 Results and Discussion

All of the models were evaluated using the minimum hamming distance, the maximum hamming distance, and the average hamming distance between predicted siRNA sequences and the given siRNA sequences. All sequences were 21 characters long. The models were also evaluated using BLEU, ROUGE-1, ROUGE-2, and ROUGE-L scores. All scores are reported in Table 1.

|  |  | **Hamming Distance** | | | **BELU** | **ROUGE** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Model** | **Min.** | **Max.** | **Avg.** | **1** | **2** | **L** |
| Sequences | Naive Bayes | 4 | 19 | 13.273 | 0.276 | 0.654 | 0.433 | 0.515 |
| Sequences | LSTM | 6 | 18 | 11.618 | 0.200 | 0.612 | 0.295 | 0.482 |
| Sequences | Transformer | 2 | 20 | 14.645 | 0.646 | 0.857 | 0.728 | 0.764 |
| Activity | Naive Bayes | 4 | 19 | 13.264 | 0.275 | 0.654 | 0.433 | 0.516 |
| Activity | LSTM | 5 | 18 | 11.736 | 0.203 | 0.626 | 0.300 | 0.483 |
| Activity | Transformer | 2 | 20 | 14.697 | 0.659 | 0.867 | 0.74 | 0.772 |
| Numerical | Naive Bayes | 4 | 19 | 13.178 | 0.271 | 0.648 | 0.428 | 0.513 |
| Numerical | LSTM | 7 | 19 | 12.39 | 0.338 | 0.67 | 0.393 | 0.569 |
| Numerical | Transformer | 0 | 18 | 8.594 | 0.493 | 0.749 | 0.633 | 0.667 |

Table 1: Scores of each model on the held-out test sets. Hamming distance scores are relative to sequence lengths of 21.

For each metric it can be observed that there are only minor relative differences in scores between the datasets. Scores in the numerical dataset tend to have the most deviance, particularly in regards to the Transformer. Given that the differences are still minor it could be taken as an indication that more features could have an impact on scores, but is most likely not the direction of research that would lead to the most improvement.



Figure 2: Comparison of Hamming Distance scores between models and datasets.

The Hamming Distance scores reported are not encouraging. The minimum distances recorded in each category are mostly reasonably low, however many of the maximums are near the entire length of the sequences. Combined with high averages clustered just above half the length of the sequences these scores indicate that all of the models are having difficulty reproducing the target siRNAs exactly.

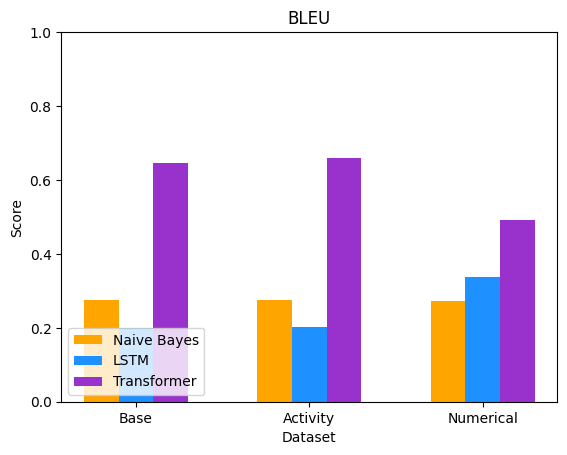


Figure 3: Comparison of BLEU scores between models and datasets.

The BLEU scores also show that the models learned poorly. As can be seen in Figure 3, most of the Naive Bayes models scored in the range of 0.2 to 0.3. Interestingly this was not the case for the Transformer models, which scored decently with scores around 0.5 and 0.65.



Figure 4: Comparison of ROUGE scores between models and datasets.

Figure 4 shows the trend of the Transformer models significantly outperforming the Naive Bayes and LSTMs continuing. While the Naive Bayes and LSTM ROUGE scores are poor to moderate in ranges of about 0.2 to 0.6, all of the Transformer ROUGE scores range from roughly 0.6 to 0.8 which is a significant improvement.

While the BLEU and ROUGE scores of the Transformer models are encouraging, when combined with the Hamming Distance scores there is a clear indication that the BLEU and ROUGE scores are likely suffering from score inflation due to the small vocabulary size. This could possibly be rectified with more data and further training. Overall the reported scores are not particularly impressive, but the significant score increase displayed by the Transformer models indicate a possible avenue of research that could yield good results.

# 4 Conclusion

The goal of this project was to build and train machine learning models that could generate siRNA sequences that would bind to a sequence of given DNA. Models of multiple architecture styles were built and trained over different versions of the initial dataset. The trained models do not perform well, but do successfully generate siRNA sequences for a given sequence of DNA. Additionally, the decent absolute ROUGE and BLUE scores as combined with the higher scores relative to the other models indicates that Transformer models have potential in this field of research.

Future work on this topic would most likely benefit from using transfer learning. Deep learning models perform better with more data to work with, and the performance of both the LSTM and Transformer suffered from low data counts. Transfer learning on similar inhibitory RNAs such as miRNA could drastically improve model performance and is worth exploring.

# 5 Author Contributions

Both authors shared research duties at the beginning, with Camrie focusing on the background biological science and Mica on what machine learning models would be appropriate. Both authors researched potential datasets. Mica handled data collection, data preprocessing, model construction, and model training. Camrie did further research into possible transfer learning datasets that were not used due to time constraints. Camrie also took the lead on creating the project presentation slides and the report.

# 

# Bibliography

Chowdhury, U. F., Sharif Shohan, M. U., Hoque, K. I., Beg, M. A., Sharif Siam, M. K., & Moni, M. A. (2021). A computational approach to design potential siRNA molecules as a prospective tool for silencing nucleocapsid phosphoprotein and surface glycoprotein gene of SARS-CoV-2. *Genomics*, *113*(1), 331–343. https://doi.org/10.1016/j.ygeno.2020.12.021

Dana, H., Mahmoodi Chalbatani, G., Mahmoodzadeh, H., Karimloo, R., Rezaiean, O., Moradzadeh, A., Mehmandoost, N., Moazzen, F., Mazraeh, A., Marmari, V., Ebrahimi, M., Menati Rashno, M., Jan Abadi, S., Gharagouzlo, E., & Mar-mari, V. (2017). Molecular Mechanisms and Biological Functions of siRNA. In *Int J Biomed Sci www.ijbs.org 48 INTERNATIONAL JOURNAL of BIOMEDICAL SCIENCE* (Vol. 13, Issue 2). www.ijbs.org

Ghosal, S., Das, S., & Chakrabarti, J. (2012). Computational Approaches for Designing Efficient and Specific siRNAs. In *Bioinformatics*. InTech. https://doi.org/10.5772/50125

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D., Labow, M., Reinhardt, M., Natt, F., & Hall, J. (2005). Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, *23*(8), 995–1001. https://doi.org/10.1038/nbt1118

Lam, J. K. W., Chow, M. Y. T., Zhang, Y., & Leung, S. W. S. (2015). siRNA versus miRNA as therapeutics for gene silencing. In *Molecular Therapy - Nucleic Acids* (Vol. 4, Issue 9, p. e252). Nature Publishing Group. https://doi.org/10.1038/mtna.2015.23

Shabalina, S. A., Spiridonov, A. N., & Ogurtsov, A. Y. (2006). Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-65

Shmushkovich, T., Monopoli, K. R., Homsy, D., Leyfer, D., Betancur-Boissel, M., Khvorova, A., & Wolfson, A. D. (2018). Functional features defining the efficacy of cholesterol-conjugated, self-deliverable, chemically modified siRNAs. *Nucleic Acids Research*, *46*(20), 10905–10916. https://doi.org/10.1093/nar/gky745

Singh, A., Trivedi, P., & Jain, N. K. (2018). Advances in siRNA delivery in cancer therapy. In *Artificial Cells, Nanomedicine and Biotechnology* (Vol. 46, Issue 2, pp. 274–283). Taylor and Francis Ltd.<https://doi.org/10.1080/21691401.2017.1307210>

Neural machine translation with a transformer and Keras&nbsp; :&nbsp; text&nbsp; :&nbsp; tensorflow. TensorFlow. (n.d.). https://www.tensorflow.org/text/tutorials/transformer

Zheng, M., Tao, W., Zou, Y., Farokhzad, O. C., & Shi, B. (2018). Nanotechnology-Based Strategies for siRNA Brain Delivery for Disease Therapy. In *Trends in Biotechnology* (Vol. 36, Issue 5, pp. 562–575). Elsevier Ltd. https://doi.org/10.1016/j.tibtech.2018.01.006

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. In *Engineering* (Vol. 6, Issue 3, pp. 275–290). Elsevier Ltd. https://doi.org/10.1016/j.eng.2019.12.014