intogenLSTM: A Sequential Machine Learning Model for Predicting Oncogenic Mutation Outcomes

Author: Krishan Bansal

In this study, we propose a Long Short-Term (LSTM) based machine learning model for predicting the consequences of cancerous mutations using prior mutation sequences. The motivation behind this work stems from the critical need for accurate prediction of mutation consequences given limited/fragmented information,which is fundamental to understanding the mechanisms underlying cancer progression and developing targeted therapeutic interventions. Our approach leverages the sequential nature of mutation data and utilizes LSTM networks, a type of recurrent neural network (RNN), to capture temporal dependencies in mutation sequences. Even when utilizing a limited amount of data and feature information for the sake of computational efficiency, performance of the model on a dataset comprising mutation triplets demonstrates promising results in terms of reliably predicting mutation consequences.

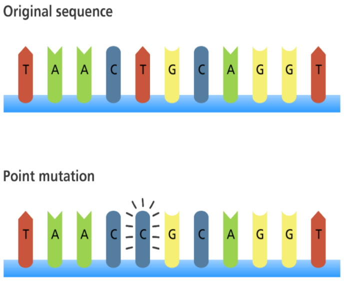
Problem Description:

Understanding the consequences of cancerous mutations is crucial for elucidating the molecular mechanisms driving cancer progression and guiding clinical decision-making. However, accurately predicting the functional impact of mutations remains a challenging task due to the complex interplay of genetic, epigenetic, and environmental factors. Traditional computational methods for predicting mutation consequences often rely on heuristic rules or biochemical properties, which may not fully capture the intricate relationships between genetic alterations and phenotypic outcomes. Furthermore, the exponential growth of genomic data presents a formidable challenge in efficiently analyzing and interpreting mutation data.

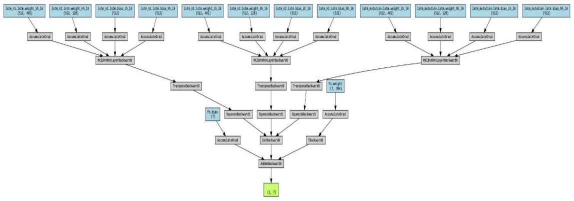
The types of mutation consequences are as follows:

1. Missense Variant: A mutation that results in a single nucleotide change, leading to the substitution of one amino acid for another in the protein sequence.
2. Nonsense Variant: A mutation that introduces a premature stop codon in the protein-coding sequence, resulting in a truncated or shortened protein.
3. Frameshift Variant: A mutation that alters the reading frame of the genetic code by inserting or deleting nucleotides, leading to a shift in the grouping of codons and often resulting in a completely different amino acid sequence downstream of the mutation.
4. Splice Variant: A mutation that affects the splicing of pre-mRNA during mRNA processing, leading to alterations in the mRNA transcript and potentially disrupting normal gene expression.
5. Synonymous Variant: A mutation that does not result in an amino acid change despite altering the nucleotide sequence, often occurring in non-coding regions or codon positions where multiple codons code for the same amino acid.
6. Inframe Deletion/Insertion: A mutation that involves the deletion or insertion of a small number of nucleotides, typically in multiples of

three, which maintains the reading frame of the genetic code and may result in the loss or gain of specific amino acids in the protein sequence.

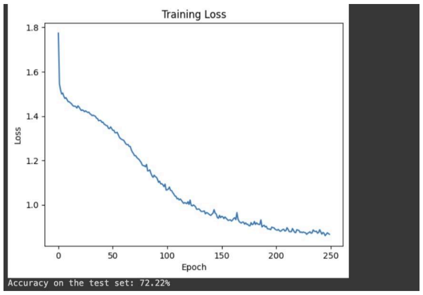


Approach:
To address the limitations of existing methods, we propose a Long Short-Term Memory (LSTM) network, which is a form of recurrent neural network which is well-suited for modeling sequential data and capturing long-range dependencies. Given a mutation triplet consisting of two neighboring mutations and a target mutation, our model learns to predict the consequence of the target mutation based on the context provided by the neighboring mutations. We formulate the mutation prediction task as a multi-class classification problem, where each class corresponds to a specific mutation consequence.



Data Description:
Our training data is derived from the intogen somatic tumor mutation database. The custom training dataset used for training comprises mutation triplets extracted from the mutation sequences of known cancer driver genes . Each mutation triplet consists of three mutations: two neighboring mutations and a target mutation. The mutations are represented as sequences of nucleotide bases, with additional metadata such as chromosome information and consequence labels. The dataset encompasses a diverse range of mutation types and consequences, providing ample variability for model training and evaluation.

Experiments and Error Analysis:
We conduct extensive experiments to evaluate the performance of our LSTM-based model on the mutation prediction task. We split the dataset into training, validation, and test sets, ensuring that each set contains a representative distribution of mutation triplets. We train the model using stochastic gradient descent with backpropagation and initially error analysis reveals that even on a limited training set loss decreases steadily over time and accuracy on the training set is well over 50%.



```
Accuracy on the test set: 72.22%
```

Discussion/Conclusion:
Our experimental results demonstrate the effectiveness of LSTM-based models in predicting mutation consequences from genomic data. 65-72% accuracy on a subset of data demonstrates the viability of using the principles presented in this study to create more complex models. One could use many more prior mutations as feature data, and also try to predict large number

outcomes. We use triplets in this study for computational efficiency, but the key takeaway is that DNA mutation outcomes appear to have a dependency on the mutation sequence of its nearest neighbors, and that these outcomes can be predicted with sequential machine learning. The model achieves competitive performance compared to existing methods, highlighting the potential of deep learning approaches in mutation analysis. Despite promising results, challenges remain in handling imbalanced datasets, mitigating overfitting, and generalizing to diverse mutation types. Future work will focus on addressing these challenges and incorporating additional features to enhance prediction accuracy and robustness.

References:

1. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
2. Liu, B., Xu, J., Lan, X., & Xu, R. (2019). Large-scale prediction of deleterious synonymous variants using deep learning. BMC bioinformatics, 20(1), 1-9.
3. Wang, M., Zhao, X., Takemoto, K., & Xu, H. (2019). Deep learning-based triplex sequence models for predicting the effects of synonymous mutations. Bioinformatics, 35(22), 4767-4775.