Predicting RNA Mutation Effects through Machine Learning of High-Throughput Ribozyme Experiments (Student Abstract)

Joseph Kitzhaber[1], Ashlyn Trapp[2], James Beck[2], Edoardo Serra[1], Francesca Spezzano[1], Eric Hayden[2], Jessica Roberts[2]

[1]Department of Computer Science, Boise State University
[2]Department of Biological Sciences, Boise State University University
{josephkitzhaber, AshlynTrapp, jimbeck, jessicaroberts1}@u.boisestate.edu
{edoardoserra, francescaspezzano, erichayden}@boisestate.edu

Abstract

The ability to study "gain of function" mutations has important implications for identifying and mitigating risks to public health and national security associated with viral infections. Numerous respiratory viruses of concern have RNA genomes (e.g., SARS and flu). These RNA genomes fold into complex structures that perform several critical functions for viruses. However, our ability to predict the functional consequence of mutations in RNA structures continues to limit our ability to predict gain of function mutations caused by altered or novel RNA structures. Biological research in this area is also limited by the considerable risk of direct experimental work with viruses. Here we used small functional RNA molecules (ribozymes) as a model system of RNA structure and function. We used combinatorial DNA synthesis to generate all of the possible individual and pairs of mutations and used high-throughput sequencing to evaluate the functional consequence of each single- and double-mutant sequence. We used this data to train a Long Short-Term Memory model. This model was also used to predict the function of sequences found in the genomes of mammals with three mutations, which were not in our training set. We found a strong prediction correlation in all of our experiments.

Introduction

Unlike human genomes that are made of DNA, many viruses have genomes made of a molecule called RNA (**r**ibo**n**ucleic **a**cid). One important property of RNA molecules is that they can form various complex shapes that are very different than the well-known double-helix of DNA. For viruses with RNA genomes, these RNA shapes can perform important functions, such as the binding of virus or human proteins that enable replication and infection. RNA molecules are formed from long chains of small molecular building blocks called nucleotides. The shape, also called structure, formed by a specific RNA molecule depends on the order of the connected nucleotides, which we call the RNA sequence. Changing the RNA sequence changes the shape, which changes the function. Random mutations that occur during the virus replication change the sequence, which can in turn change the shape and function of the RNA. Some of these mutations, termed "gain of function", would make the virus better at replicating or infecting, which makes them more harmful. Unfortunately, our inability to predict which sequence changes will change RNA functions limits our ability to predict gain of function mutations. The ability to predict gain of function mutations in RNA could help us prepare and respond to viral epidemics and pandemics.

Self-cleaving ribozymes are small functional RNA molecules that can be found in the genomes of all living organisms. Ribozymes are a good model of sequence to function relationships in RNA because changing the sequence of a ribozyme can change the function, and this can be studied safely and easily in the lab. "Gain of function mutations" enhance the ribozyme function, but the ribozymes are not infectious or harmful to humans. In addition, it is easy to make changes to the RNA sequence in the lab. Recent experimental advancements have made it possible to study millions of nucleotide changes to a ribozyme sequence all at once, providing rich data sets for learning and predicting which nucleotide changes result in gain of function mutations in this model RNA system.

Here we use such high-throughput sequence-function data from a ribozyme called CPEB3. This ribozyme was originally found in the human genome but now is known to have been present in the ancient genomes of the earliest mammals. The ribozyme is found in the genomes of all mammals but with some differences in the nucleotide sequences. This provides a good model system for predicting which nucleotide changes in mammals enhanced ribozyme function. Toward this goal, we first generated all the possible nucleotide changes to the most ancient CPEB3 ribozyme, all the possible pairs of nucleotide changes, and evaluated the ribozyme function for all these sequences ($\sim$20,000 sequences). We used a portion of this data as a training set for a Long Short-Term Memory model (LSTM). We used this model to predict the function of the withheld training data and then to predict the function of sequences with three mutations that were not in our training data but are found in the genomes of other mammals. We analyzed the accuracy of our predictive model using correlation between the predicted ribozyme activity and the lab determined activity.

Previous works focusing on functional prediction aim to predict mutated sequences with the same number of mutations as in the training data (Zhang et al. 2020; Schmidt and

	R^2	*RMSE*	*PCC*
Experiment 1: Replicate 1	0.736	0.152	0.858
Experiment 1: Replicate 2	0.841	0.121	0.918
Experiment 1: Replicate 3	0.848	0.109	0.921
Experiment 2: Replicate 1	0.648	0.180	0.892
Experiment 2: Replicate 2	0.639	0.174	0.924
Experiment 2: Replicate 3	0.800	0.143	0.918

Table 1: Prediction scores for Experiments 1 and 2.

Smolke 2021; Calonaci et al. 2020). Most of these models enhance structural homology to predict, i.e., estimating how similar is the sequence to predict with the one in input. To the best of our knowledge, this is the first time that a model trained on single and double mutations is effective to predict sequences with a larger number of mutations. This is possible because our LSTM model does not use homology but learns specific small patterns in the sequence that can also generalize to sequences with a higher number of mutations.

Methods

Data Generation. The nucleotide changes to the CPEB3 ribozyme were made and studied through molecular biology techniques as follows. The RNA molecules were made in the lab through a process called in vitro transcription, where a protein is used to make multiple RNA copies of DNA "templates". Ribozymes are unique RNA sequences that cleave their own sequence while they are being made (self-cleaving). The DNA templates used in this study were actually a collection of numerous different DNA sequences, and the RNA that was made included lots of copies of every possible individual and pair of nucleotide changes of the CPEB3 ribozyme. Some of these nucleotide changes were expected to enhance the ribozyme function, such that more of the molecules will cleave while they are being made. Other nucleotide changes were expected to diminish the ribozyme activity, such that less of these molecules will cleave while being made. All the different RNA sequences were made simultaneously so that the amount that they cleave could be directly compared. The mount cleaved was determined by sequencing all the RNA molecules (RNA-seq). This sequencing data reports on both the nucleotide changes that were present in a specific molecule and whether or not that specific molecule was cleaved. Because each nucleotide change was observed multiple times, counting the number of cleaved and not cleaved molecules was used to determine the function of each sequence. Specifically, the function was defined as the fraction cleaved F = count cleaved/count total.

The lab experiments were replicated three times because the molecular biology methods used involve several stochastic processes. Each replicate was used separately to train and test the prediction model and contains 207 sequences with one mutation and 21,114 sequences with double mutation.

Prediction Model. We built a machine learning model to predict how nucleotide changes will change the ribozyme function (fraction cleaved). Each dataset contains nucleotide sequences and the associated fraction cleaved determined experimentally, which we used as ground truth. Because the fraction cleaved is a real number in [0,1], we addressed the problem as a regression task. Given that the input is a sequence of nucleotides, we used a Long Short-Term Memory modelto predict the fraction cleaved. We processed the input sequence of nucleotides as a text sequence. We used the root mean square error (RMSE) as the loss function. We used R^2 score, RMSE, and Pearson correlation coefficient (PCC) to measure the model performances of the proposed model.

Experiments and Results

We conducted two types of experiments. In the first experiment (Experiment 1), we trained and predicted on individual and pairs of nucleotide mutations. We performed 10-fold cross-validation for each of the Replicates. Results are shown in Table 1 (first three rows). In the second experiment (Experiment 2), we wanted to see if a model learned on one and two mutations was able to predict the function of sequences with a higher number of mutations (three in the considered case). Hence, we used all the Replicate data as training set and data consisting of sequences found in the genomes of mammals with three mutations, which were not in our training set (518 sequences). Results are shown in Table 2 (second three rows). We found that there was a strong prediction correlation in all models. For both experiments 1 and 2, the Pearson Correlation Coefficient on Replicates 2 and 3 was $>$0.9, and the R^2 ranged from 0.64 to 0.84. These results indicate that tractable high-throughput experiments in the lab can be used to determine the effect of combinations of mutations found to evolve in nature. Results were slightly worse on Replicate 1. Because of variation in the total number of copies for all mutational variants within each replicate, the average number of identified copies per variant is different. Additionally, because this difference is not uniformly distributed across all mutational variants, some variant's counts may be less accurate due to low copy volume. Consequently, we should not expect that each replicate will be of equal utility for downstream predictions. Overall, these preliminary results are promising for predicting "gain of function" mutations in other settings, such as RNA viruses, which could guide RNA vaccine production.

Acknowledgments

This research has been sponsored by the National Science Foundation under award #1950599.

References

Calonaci, N.; Jones, A.; Cuturello, F.; Sattler, M.; and Bussi, G. 2020. Machine learning a model for RNA structure prediction. *NAR genomics and bioinformatics*, 2(4): lqaa090.

Schmidt, C. M.; and Smolke, C. D. 2021. A convolutional neural network for the prediction and forward design of ribozyme-based gene-control elements. *Elife*, 10: e59697.

Zhang, Z.; Xiong, P.; Zhang, T.; Wang, J.; Zhan, J.; and Zhou, Y. 2020. Accurate inference of the full base-pairing structure of RNA by deep mutational scanning and covariation-induced deviation of activity. *Nucleic acids research*, 48(3): 1451–1465.