Vattikonda, N., Aha, D.W., Jackson, A., Leary, D. (2017). Validating and Improving Prokaryotic Promoter Prediction Models (NCARAI Technical Note AIC-17-5510-101). Washington, D.C.: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence.

Validating and Improving Prokaryotic Promoter Prediction Models

Niharika Vattikonda[1], David W. Aha[1], Amina Jackson[2], & Dasha Leary[2]
[1]Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory (Code 5514); Washington, DC 20375
[2]Navy Center for Biomolecular Science and Engineering
Naval Research Laboratory (Code 6910); Washington, DC 20375
niharika.vattikonda@gmail.com | {*first.last*}@nrl.navy.mil

18 August 2017

Abstract

Promoters are essential in the process of transcribing DNA to mRNA, which is needed for protein creation. Certain nucleotide sequences throughout the genome are promoters because the structure created by the nucleotides in the promoter sequence allows the RNA polymerase to bind to the promoter. In an earlier approach, a recurrent neural network was used to predict the probability that a sequence of nucleotides was a promoter based on models of *Escherichia coli* and *Bacillus subtilis*. Unfortunately, it oversimplified the complexity of the structure generated by nucleotide interactions and searched for recurring patterns within sequences, which did not consider that many promoters tend to be gene-specific. This resulted in a high rate of false positive identifications. Therefore, we developed a new approach to decrease the false identification rate. It relies on details about the characterization of promoters that are not directly related to the nucleotide sequence (e.g., the relatively higher frequency of promoters compared to other sequences within a genome). Although we have not yet analyzed relative frequency, retraining promoter models decreased the false positive rate and decreased the false identification rate.

1. Introduction

Promoters are typically gene-specific regions of a genome that are key in starting and regulating the process of gene transcription. Essentially, the nucleotide sequence of a promoter causes that section of the genome to have a specific structural shape that the RNA polymerase, the agent responsible for the actual process of transcription, can bind to. While certain characteristics and nucleotide patterns are found in many promoters, the gene-specific nature of promoter sequences complicates identification. Being able to identify promoters accurately in a genome that has not yet been fully documented is vital to being able to identify where genes start and could be useful in genetic modification and regulation of gene transcription. In the original approach, the genome of the organism being studied was divided into 80-nucleotide sequences, and the nucleotide sequences were converted to a string of numbers (nucleotides were translated to numbers as follows: adenine = 1, thymine = 2, cytosine = 3, guanine = 4) and then input to a recurrent neural network; the output was a probability that the input sequence was a promoter

(expressed as a decimal value from 0 to 1). The models were originally trained on *Escherichia coli* and *Bacillus subtilis* separately but blended models were used later because training a model on a single organism caused overfitting to the training organism and failed to predict promoters accurately for other organisms.

We decided that the original approach was too one-dimensional and focused on merely recognizing patterns within sequences of text, which did not accurately demonstrate the complexity of promoter sequences. Rather than being merely a string of nucleotides, the more important function of a promoter is to have a sequence of nucleotides such that the structure created by that sequence can bind to the RNA polymerase. Moreover, this original approach failed to account for the frequency of certain similar or duplicated nucleotide sequences throughout the genome, and since promoters typically tend to occur throughout the genome, this approach was missing a key property of a promoter. Therefore, a new method to validate and improve the models was designed to add more complexity to the results generated by the models and to remove any confounding factors that could cause the model to train on the wrong nucleotide patterns within the sequences. The techniques of this approach to validating promoter prediction data to improve the models were creating a binary matrix based on the positions of unique sequences (to determine if matches and duplicates are present) and removing the DNA start codon and open reading frames from predicted promoters to retrain the models.

2. Binary Matrix of Sequence-Position Data

The purpose of this binary matrix of sequence and position data is to determine which sequences appear most frequently, as promoter sequences are likely to appear multiple times throughout a genome. This relatively higher frequency of a promoter compared to a gene or other segment of the genome is due to the fact that a single promoter can activate the transcription of multiple different genes spaced out across the genome. In the promoter prediction results, each sequence of 80 nucleotides was displayed in a row with its position on the genome and the probability that the sequence was a promoter (assigned as a decimal from 0 to 1 by the model). For this binary matrix, duplicate sequences were first removed from the set, and unique sequences were used as the rows. Positions were used as the headers of columns, and for each sequence, if the sequence indicated by the row appeared at the position specified by the column header value, the entry was given a value of 1. After the results were processed to assign 1s to the binary matrix, unfilled entries were filled with the value 0, indicating that the sequence was not found at the given position. The binary matrix was produced successfully when tested on 10, 20, 50, 100, and 200 sequences, and the full binary matrix (with approximately 89,000 sequences) will be created for *Bacillus subtilis,* one of the organisms used to train the models. The binary matrix format also solves some of the data formatting issues with bioinformatics tools, since the binary matrix filters the data in such a way that there is less additional processing required to input the matrix into tools such as GenePattern. However, the full binary matrix has not been created and clustered yet, primarily due to adjustments that needed to be made to the models due to discrepancies that were found while manually reviewing predicted promoter sequences, as detailed below in Section 3.

3. Removal of the DNA Start Codon and Open Reading Frames from Predicted Promoter Sequences to Retrain Models

While manually reviewing the sequences that the model predicted as promoters, we discovered that all the predicted promoter sequences included the 3-nucleotide DNA sequence "ATG" at the end of the sequence. However, the sequence "ATG" does not indicate a promoter; rather, "ATG" comprises the start codon, which is where transcription of the gene begin, and is unlike the promoter, which is where the RNA polymerase binds before transcription begins. While the promoter sequence is removed once the gene has gone through transcription, the start codon is still part of the mRNA sequence that results from transcription and is carried over to the next process of protein creation: mRNA translation. It was determined that the model had likely been picking up "ATG" as a characteristic of a potential promoter sequence (since many promoters end right before the start codon), which would have led to incorrect predictions and was a potential explanation for the rate of false positive identification. Therefore, to remove this confounding variable, the last three characters ("ATG") of every predicted promoter sequence were removed in the results file.

Furthermore, after reviewing some of the predicted promoter sequences manually, it was noted that many of the sequences had open reading frames contained within the predicted promoter sequence. An open reading frame (ORF) is any part of a reading frame (the current part of the genome being processed) that could be translated, meaning that within the current reading frame, there is a DNA sequence that begins with the start codon ("ATG") and ends with a stop codon, which could be one of three potential codons: "TAA," "TAG," and "TGA." Because a promoter is the site of attachment for the RNA polymerase, the promoter sequence does not code for a protein; leaving open reading frames in the predicted promoters could cause the model to mistakenly include characteristics of ORFs as characteristics of promoters, which would contribute to the inaccuracy of the model. While processing the results file yet again, the 200-nucleotide predicted promoters were cut into 80-nucleotide sequences, each staggered by a single nucleotide. In processing the 80-nucleotide sequences that were cut from the original predicted promoter sequences, any ORFs (including the start and stop codons) were removed from the sequence, although the prediction value assigned by the model was not modified. Table 1 shows the original confusion matrix for *Vibrio natriegens* with a model that was only trained on *Escherichia coli* and demonstrates a 49.8% accuracy, since the model essentially predicted all sequences as promoters; however, when the model was trained on a blended data set (of *Bacillus subtilis* and *Vibrio natriegens*), the model stopped predicting every sequence as a promoter, and once the predicted promoters were adjusted by the process described above, the model demonstrated 86.3% as seen in Table 2, with a false negative rate of just 1.2%.

	Predicted = No	Predicted = Yes	
Actual = No	0.0%	50.2%	Total = 50.2%
Actual = Yes	0.0%	49.8%	Total = 49.8%
	Total = 0.0%	Total = 49.8%	

Table 1. Confusion matrix for *Vibrio natriegens* with the original model that was only trained on *Escherichia coli* that attained an accuracy of 49.8% and a 50.2% false positive identification rate.

	Predicted = No	Predicted = Yes	
Actual = No	34.6%	12.5%	Total = 47.1%
Actual = Yes	1.2%	51.7%	Total = 52.9%
	Total = 35.8%	Total = 64.2%	

Table 2. Confusion matrix for *Vibrio natriegens* after the model was retrained using the adjusted promoters (without the start codon or any open reading frame) that attained an accuracy of 86.3%.

4. Conclusion and Next Steps

Based on the results of the model that was retrained using the adjusted predicted promoters (without the start codon or any open reading frames), retraining the models on adjusted predicted promoters improves the accuracy of the models, as demonstrated by the decrease in false negatives and false positives in the two confusion matrices (Tables 1 and 2). Currently, this

adjusted predicted promoter approach increases the accuracy of the model for *Vibrio natriegens*, which is not one of the organisms that the model was originally trained on. The next step would be to evaluate if this approach is beneficial for *Escherichia coli* or *Bacillus subtilis*, the two organisms that the model was originally trained on, and then to test the adjusted predicted promoter approach on different prokaryotes that already have some experimentally determined promoters, so that there is a baseline to compare predictions to the actual promoters.

Once the binary matrix is processed for *Bacillus subtilis*, the data will be much easier to process using bioinformatics clustering tools such as GenePattern. Once the clustering algorithms have been run on the binary matrix, a Rand index implementation should be run to evaluate the quality of the clusters by penalizing false negative and false positive results; originally, some predicted promoter sequences seemed to be false positives (since there were more positive results than a plausible number of promoters) in the clusters, so the Rand index implementation would help identify these issues. After the quality of the clusters is evaluated, promoters that appear most frequently, as determined by the clustering process, will be checked against the promoters that are predicted by the model in order to determine whether the model is generating valid prediction data. While this initial test will be on *Bacillus subtilis*, the next organism to be processed should be *Escherichia coli*, since that is the other organism that the models were initially trained on.

Acknowledgements

Niharika Vattikonda is a 12th grade student at Thomas Jefferson High School for Science and Technology in Alexandria, Virginia. She thanks the Naval Research Laboratory for allowing her to participate as a Student Volunteer during the summer of 2017. She also thanks David Aha and Dasha Leary for giving her the opportunity to pursue research at the intersection of genomics and computer science and Amina Jackson for providing assistance with developing the approach that Niharika focused on improving during her internship this summer.